

# **Alaska Comprehensive System of Student Assessment**

## **2003 Spring Technical Report Benchmark Assessments and High School Graduation Qualifying Exam**

**Submitted to  
The Alaska Department of Education & Early Development**



**CTB/McGraw-Hill  
20 Ryan Ranch Road  
Monterey, California 93940**



This technical report provides descriptions of the testing procedures used for the analysis of the operational form of the Benchmark Assessments (Benchmark) administered at Grades 3, 6 and 8, and the Alaska High School Graduation Qualifying Exam (HSGQE) initially administered at Grade 10 in the spring of 2002. Both the Benchmark and HSGQE covered three content areas: Reading, Mathematics, and Writing. The intended audience of this report is the Alaska Department of Education & Early Development (ADEED), the Technical Advisory Committee, and CTB/McGraw-Hill.

## Table of Contents

<b>TABLE OF CONTENTS .....</b>	<b>4</b>
<b>INTRODUCTION .....</b>	<b>7</b>
<b>VALIDITY .....</b>	<b>8</b>
CONTENT VALIDITY .....	8
CONSTRUCT VALIDITY .....	9
TEST CONTENT ALIGNMENT .....	14
<b>ITEM SELECTION .....</b>	<b>16</b>
ITEM SELECTION PROGRAM .....	16
ITEM SELECTION PROCESS .....	16
ITEM SELECTION RESULTS .....	17
<b>INDIVIDUAL ITEM ANALYSES .....</b>	<b>19</b>
<b>DESCRIPTIVE STATISTICS AND RELIABILITY .....</b>	<b>34</b>
PERCENTAGE OF STUDENTS IN EACH PROFICIENT CATEGORY .....	36
STANDARD ERROR OF MEASUREMENT .....	36
INTER-RATER RELIABILITY .....	40
<b>CALIBRATION AND EQUATING .....</b>	<b>45</b>
CALIBRATION .....	46
3PL/2PPC MODELS .....	46
ITEM FIT AND NONCONVERGENCE .....	48
DIMENSIONALITY .....	49
SETTING THE SCALE UNITS AND VALUES FOR THE LOSS AND HOSS .....	50
EQUATING: THE 2003 OPERATIONAL TEST SCALE .....	50
BIAS STUDIES .....	51
BIAS REVIEWS .....	51
DIFFERENTIAL ITEM FUNCTIONING .....	53
<i>Linn and Harnisch</i> .....	54
<i>Standardized Mean Difference</i> .....	57
<b>ALASKA PERFORMANCE INDEX .....</b>	<b>62</b>
API CUTPOINT .....	63
HSGQE FIELD TEST .....	67
<b>APPENDIX A: FIT MEASUREMENT: A GENERALIZATION OF Q1 .....</b>	<b>70</b>
<b>APPENDIX B: ALASKA PERFORMANCE INDEX PROCEDURE .....</b>	<b>72</b>
<b>REFERENCES .....</b>	<b>74</b>

## List of Tables

TABLE 1 – NUMBER OF OPERATIONAL ITEMS ADMINISTERED BY ITEM TYPE FOR OPERATIONAL TESTS.....	7
TABLE 2 – ITEM BREAKOUT BY CONTENT PERFORMANCE STANDARD – GRADE 3 .....	10
TABLE 3 – ITEM BREAKOUT BY CONTENT PERFORMANCE STANDARD – GRADE 6 .....	11
TABLE 4 – ITEM BREAKOUT BY CONTENT PERFORMANCE STANDARD – GRADE 8 .....	12
TABLE 5 – ITEM BREAKOUT BY CONTENT PERFORMANCE STANDARD – HSGQE .....	13
TABLE 6 – CONTENT PROPORTIONALITY – HSGQE SPRING 2003 READING .....	14
TABLE 7 – CONTENT PROPORTIONALITY – HSGQE SPRING 2003 MATHEMATICS .....	15
TABLE 8 – CONTENT PROPORTIONALITY – HSGQE SPRING 2003 WRITING .....	15
TABLE 9 – ITEM DIFFICULTY FREQUENCY DISTRIBUTION (HSGQE) .....	18
TABLE 10 – ITEM DIFFICULTY FREQUENCY DISTRIBUTION BY SCORE POINTS FOR HSGQE SPRING 2003.....	19
TABLE 11 – BENCHMARK 1 READING ITEM STATISTICS .....	20
TABLE 12 – BENCHMARK 1 MATHEMATICS ITEM STATISTICS .....	21
TABLE 13 – BENCHMARK 1 WRITING ITEM STATISTICS .....	22
TABLE 14 – BENCHMARK 2 READING ITEM STATISTICS .....	23
TABLE 15 – BENCHMARK 2 MATHEMATICS ITEM STATISTICS .....	24
TABLE 16 – BENCHMARK 2 WRITING ITEM STATISTICS .....	25
TABLE 17 – BENCHMARK 3 READING ITEM STATISTICS .....	26
TABLE 18 – BENCHMARK 3 MATHEMATICS ITEM STATISTICS .....	27
TABLE 19 – BENCHMARK 3 WRITING ITEM STATISTICS .....	28
TABLE 20 – HSGQE READING ITEM STATISTICS .....	29
TABLE 21 – HSGQE MATHEMATICS ITEM STATISTICS.....	31
TABLE 22 – HSGQE WRITING ITEM STATISTICS .....	33
TABLE 23 – DESCRIPTIVE STATISTICS AND RELIABILITY.....	35
TABLE 24 – PERCENTAGE OF STUDENTS IN EACH PROFICIENT CATEGORY FOR BENCHMARK.....	36
TABLE 25 – PERCENTAGE OF STUDENTS IN EACH PROFICIENT CATEGORY FOR HSGQE.....	36
TABLE 26 – SCALE SCORE SEM’S AND 80% CONFIDENCE INTERVALS.....	38
TABLE 27 – RAW SCORE BASED SEM COMPARISON TO TERRANOVA FOR 2002 .....	39
TABLE 28 – BENCHMARK 1 MATHEMATICS RATER ANALYSES.....	40
TABLE 29 – BENCHMARK 1 READING RATER ANALYSES.....	41
TABLE 30 – BENCHMARK 1 WRITING RATER ANALYSES.....	41
TABLE 31 – BENCHMARK 2 MATHEMATICS RATER ANALYSES.....	41
TABLE 32 – BENCHMARK 2 READING RATER ANALYSES.....	42
TABLE 33 – BENCHMARK 2 WRITING RATER ANALYSES.....	42
TABLE 34 – BENCHMARK 3 MATHEMATICS RATER ANALYSES.....	42
TABLE 35 – BENCHMARK 3 READING RATER ANALYSES.....	43
TABLE 36 – BENCHMARK 3 WRITING RATER ANALYSES.....	43
TABLE 37 – HSGQE MATHEMATICS RATER ANALYSES.....	43
TABLE 38 – HSGQE READING RATER ANALYSES.....	44
TABLE 39 – HSGQE WRITING RATER ANALYSES.....	44
TABLE 40 – EXACT AGREEMENT RATES BY GRADE/CONTENT AREA.....	45
TABLE 41 – AVERAGE AGREEMENT RATES BY GRADE/CONTENT AREA.....	45
TABLE 42 – SUMMARY OF CALIBRATION RESULTS.....	48
TABLE 43 – NUMBER OF MISFIT ITEMS.....	49
TABLE 44 – ITEM PAIR DEPENDENCE BY GRADE/CONTENT AREA.....	50
TABLE 45 – SUMMARY OF EQUATED ITEM PARAMETERS FOR HSGQE .....	51
TABLE 46 – ALASKAN REGIONAL CLASSIFICATIONS.....	52
TABLE 47 – ALASKAN CLASSIFICATION AND CRITERIA FOR COMMUNITY TYPE.....	52
TABLE 48 – NUMBER OF SIGNIFICANT DIF – LINN-HARNISCH STATISTICS .....	56
TABLE 49 – DIF RATING CRITERIA.....	57
TABLE 50 – SUMMARY OF MEASURED DIF – ITEM SUMMARY TABLE – HSGQE .....	58
TABLE 51 – SUMMARY OF MEASURED DIF – ITEM SUMMARY TABLE – BENCHMARK 1 .....	59
TABLE 52 – SUMMARY OF MEASURED DIF – ITEM SUMMARY TABLE – BENCHMARK 2 .....	60
TABLE 53 – SUMMARY OF MEASURED DIF – ITEM SUMMARY TABLE – BENCHMARK 3 .....	61
TABLE 54 – BENCHMARK 1 API CUTPOINTS.....	63
TABLE 55 – BENCHMARK 2 API CUTPOINTS.....	64
TABLE 56 – BENCHMARK 3 API CUTPOINTS.....	65
TABLE 57 – HSGQE API CUTPOINTS.....	66

<b>TABLE 58 – NUMBER OF FIELD TEST ITEMS ADMINISTERED BY ITEM TYPE FOR HSGQE .....</b>	<b>67</b>
<b>TABLE 59 – FIELD TEST P – VALUE RESULTS .....</b>	<b>68</b>
<b>TABLE 60 – SUMMARY OF CALIBRATION RESULTS ON OPERATIONAL AND FIELD TEST ITEMS – HSGQE .....</b>	<b>68</b>
<b>TABLE 61 – NUMBER OF MISFIT FIELD TEST ITEMS – HSGQE .....</b>	<b>69</b>

## Introduction

This technical report will provide descriptions of the testing procedures used for the analysis of the operational form of the Benchmark Assessments (Benchmark) administered at Grades 3, 6 and 8, and the Alaska High School Graduation Qualifying Exam (HSGQE) administered at Grade 10 in the spring of 2003. The content areas covered were Reading, Writing, and Mathematics at all grade levels. This exam was administered to the students of the state of Alaska in the spring of 2003.

Table 1 shows the content areas and the number of operational items of a given type administered at each grade for the operational test.

**Table 1 – Number of Operational Items Administered by Item Type for Operational Tests**

Grade	Content Area	MC	SCR 1 pt	SCR 2 pt	SCR 3 pt	ECR 4 pt	ECR 5 pt	ECR 6 pt	Total	Score Points
3	Reading	30	2	3	–	1	–	–	36	42
	Mathematics	30	–	5	–	1	–	–	36	44
	Writing	30	1	–	–	3	–	2	36	55
6	Reading	30	2	3	1	–	–	–	36	41
	Mathematics	28	–	6	–	2	–	–	36	48
	Writing	29	1	–	–	4	–	2	36	58
8	Reading	30	2	3	1	–	–	–	36	41
	Mathematics	29	–	6	–	1	–	–	36	45
	Writing	30	–	–	–	4	–	2	36	58
HSGQE	Reading	36	–	8	5	1	–	–	50	71
	Mathematics	50	–	6	–	2	–	–	58	70
	Writing	26	1	1	–	6	–	1	35	59*

\* This score point is unweighted

Note: MC = Multiple Choice; SCR = Short Constructed Response; ECR = Extended Constructed Response

This report opens with the “Validity” section which describes the procedures followed to develop the assessments and establish their content validity. The sub-sections “Content Validity” and “Construct Validity” address initial conclusions of what the test scores mean and what types of inferences they support.

The section following “Validity” is titled “Item Selection.” This section explains the program, process, and results associated with the selection of items for the 2003 Benchmark and HSGQE tests.

The raw score and scale score descriptive statistics are then discussed and evaluated in the “Descriptive Statistics and Reliability” sections, which also include descriptions of item difficulty, standard error of measurement, and an evaluation of the reliability of the hand-scoring process through the Inter-Rater Reliability study.

The section titled “Calibration and Equating” documents the 2003 calibration procedure and the additional steps taken to ensure that the parameter estimates of the 2003 operational items were placed on the common scale.

The possibility that the Benchmark items and the HSGQE items would function differently for groups within the testing populations was investigated. The committee review, statistical procedures, and results are described in the “Bias Studies” section. The Alaska Performance Index (API) and the procedures used to establish the index are discussed in this report.

## Validity

### Content Validity

The American Psychological Association (APA) *Standards for Educational and Psychological Testing* (1999) addressed the concept of validity in testing:

Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process for accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself (p.9).

Generally, achievement tests are used for either (1) making predictions about students, or (2) describing students' performance (Mehrens & Lehmann, 1991). The second purpose is most relevant for the HSGQE and the Benchmark. The purpose of the Benchmark and the HSGQE is to document student performance in the areas of Reading, Mathematics, and Writing as defined by the 1999 Alaska Performance Standards (APS). To ensure that test scores allow interpretations appropriate for this purpose, the content of the test must be carefully matched to the specified standards. Evidence of content-related validity is of primary importance in the Benchmark and the HSGQE. The 1999 APA Standards state:

Content-related evidence of validity is a central concern during test development. Expert professional judgment should play an integral part in developing the definition of what is to be measured, such as describing the universe of the content, generating or selecting the content sample, and specifying the item format and scoring system (p.12).

Unfortunately, as Brown (1976, p. 123) has noted, there are no well-established or satisfactory numerical indices to indicate the match of item content to objectives, test content to curricular materials, or skills measured to skills taught. Reliance must be placed on human judgment. Accordingly, the content validity of the Benchmark and the HSGQE was determined by judging the extent to which test construction plans and procedures could reasonably be assumed to ensure validity. The general procedures used in test development were as follows:

1. The Alaska Performance Standards were developed with the involvement of instructional specialists.
2. The standards and skills were deemed acceptable. Educators and citizens were involved in this process.
3. Item specifications were written for each of the Alaska Performance Standards.
4. Test items were selected and/or revised according to the guidelines provided by the item specifications.
5. Instructional specialists and experienced teachers reviewed the draft items, recommending revisions when necessary.



6. The test items were subjected to final editing and prepared for inclusion in the field test forms.

Drafts of performance standards were developed by the Alaska Department of Education & Early Development (ADEED). These performance standards were then extensively reviewed and critiqued by district educators and citizens throughout the state. The final step in the process was the adoption of the Alaska Performance Standards (APS) by the State Board of Education after a public hearing.

Steps 3 through 6 of the above procedures were accomplished through a 1999 test development contract with CTB/McGraw-Hill. The test development contract has proceeded in the following manner. The specifications to be included in the tests were written for each performance standard. Draft items were written according to the test specifications, then subjected to several content reviews. All items were then reviewed for cultural, ethnic, language, and gender bias. In addition, all items were reviewed for issues of general concern to Alaska citizens. Constructed response items were pilot tested in local classrooms. Next, items were field tested to provide estimates of Alaska students' performance on the items. Finally, the operational test forms were administered in 2000. The Spring 2000 administration of Form A provides the baseline scale for the subsequent Benchmark and the HSGQE assessments. This administration also produces results of Alaska students' performance on the operational forms. The operational test results for the Spring 2003 administration are reported in this document.

### Construct Validity

Construct validity is defined as the degree to which a test measures what it was designed to measure. For instance, do the test scores represent the knowledge and performance the test was designed to measure? What kinds of inferences can be made based on the students' scores on the exam? Construct validity is the central concept underlying the Benchmark and the HSGQE validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity. For example, to demonstrate comprehensiveness, the Benchmark and the HSGQE must contain items that represent essential instructional objectives. Additionally, patterns of correlations among the content areas should demonstrate convergent and discriminate validity. That is, tests designed to measure similar skills should correlate more than tests designed to measure distinctly different skills.

The presence of these two types of evidence protects against construct under-representation and construct-irrelevant variance in a test (Cook & Campbell, 1979). The threat posed to construct validity by construct under-representation is that the test is too narrow and fails to include important aspects of the construct. The threat of construct-irrelevant variance is that the test contains excess reliable variance that is irrelevant to the interpreted construct.

Tables 2–5 present the percentage of items tested for each content performance standard for the Benchmark exams (grades 3, 6, and 8) and the HSGQE in the content areas of Reading, Mathematics, and Writing. More than 70% of the total items for each content area are multiple-choice items.

**Table 2 – Item Breakout by Content Performance Standard – Grade 3**

Content Area/Standard		Percent of Items		
Content Standard	Title	Multiple Choice	Constructed Response	Total
<b>Reading</b>				
R1.01	Use phonics; read words	10.8	0.0	10.8
R1.02	Comprehend literal meaning	16.3	2.7	19.0
R1.04	Retell or restate information	8.1	2.7	10.8
R1.05	Identify main idea	10.8	2.7	13.5
R1.06	Follow simple directions	8.1	2.7	10.8
R1.07	Identify forms of texts	10.8	0.0	10.8
R1.08	Identify basic story elements	5.4	5.4	10.8
R1.10	Connections	13.5	0.0	13.5
	<b>*Total</b>	<b>83.8</b>	<b>16.2</b>	<b>100.0</b>
<b>Mathematics</b>				
B/C/D	Prob. Solve/Comm/Reasoning	0.0	13.9	13.9
M.A.1	Numeration	16.6	2.8	19.4
M.A.2	Measurement	13.9	2.8	16.7
M.A.3	Estimation & Computation	19.4	2.8	22.2
M.A.4	Functions & Relationships	11.1	2.8	13.9
M.A.5	Geometry	13.9	2.8	16.7
M.A.6	Statistics/Probability	8.3	2.8	11.1
	<b>**Total</b>	<b>83.2</b>	<b>16.8</b>	<b>100.0</b>
<b>Writing</b>				
W1.1/1.2	Write short story	30.6	8.3	38.9
W1.3	Proofread writing	33.3	5.6	38.9
W1.4	Revise writing for clarity	19.4	2.8	22.2
	<b>Total</b>	<b>83.3</b>	<b>16.7</b>	<b>100.0</b>

A single Reading item is accounted for twice in two separate Reading Content Standards.

There are 36 items in Reading, Grade 3. Item #26 counts toward R1.02 and R1.08.

\* Therefore, the percentages are based on the total of 37 items.

\*\* The mathematics total does not include the B/C/D objective.

**Table 3 – Item Breakout by Content Performance Standard – Grade 6**

Content Area/Standard		Percent of Items		
Content Standard	Title	Multiple Choice	Constructed Response	Total
<b>Reading</b>				
R2.01	Use phonics; read words	8.3	2.8	11.1
R2.02	Infer meaning; identify themes	8.3	2.8	11.1
R2.04	Retell or summarize info.	11.1	0.0	11.1
R2.05	Connect main ideas	8.3	2.8	11.1
R2.06	Follow multi-step directions	13.9	2.8	16.7
R2.07	Describe forms of texts	11.1	0.0	11.1
R2.08	Define basic story elements	11.1	2.8	13.9
R2.09	Differentiate fact from fiction	11.1	2.8	13.9
	Total	83.2	16.8	100.0
<b>Mathematics</b>				
B/C/D	Prob. Solve/Comm/Reasoning	0.0	16.7	16.7
M.A.1	Numeration	13.9	2.8	16.7
M.A.2	Measurement	16.6	2.8	19.4
M.A.3	Estimation & Computation	16.6	2.8	19.4
M.A.4	Functions & Relationships	13.9	2.8	16.7
M.A.5	Geometry	13.9	2.8	16.7
M.A.6	Statistics/Probability	2.8	8.3	11.1
	*Total	77.7	22.3	100.0
<b>Writing</b>				
W2.1/2.2	Write about a topic	0.0	13.9	13.9
W2.3	Proofread writing	41.7	2.8	44.5
W2.4	Revise writing for support	38.8	2.8	41.6
	Total	80.5	19.5	100.0

\* The mathematics total does not include the B/C/D objective.

**Table 4 – Item Breakout by Content Performance Standard – Grade 8**

Content Area/Standard		Percent of Items		
Content Standard	Title	Multiple Choice	Constructed Response	Total
<b>Reading</b>				
R3.01	Read unfamiliar words	8.3	2.8	11.1
R3.10	Support understanding of theme	8.3	5.6	13.9
R3.04	Restate or summarize	11.1	5.6	16.7
R3.05	Assess support of main idea	13.9	0.0	13.9
R3.06	Follow multi-step directions	11.1	0.0	11.1
R3.07	Identify rules of forms of texts	11.1	0.0	11.1
R3.08	Analyze basic story elements	8.3	2.8	11.1
R3.09	Analyze authors purpose	11.1	0.0	11.1
	Total	83.2	16.8	100.0
<b>Mathematics</b>				
B/C/D	Prob. Solve/Comm/Reasoning	0.0	16.7	16.7
M.A.1	Numeration	11.1	2.8	13.9
M.A.2	Measurement	13.9	2.8	16.7
M.A.3	Estimation & Computation	13.9	2.8	16.7
M.A.4	Functions & Relationships	19.3	2.8	22.1
M.A.5	Geometry	13.9	2.8	16.7
M.A.6	Statistics/Probability	8.3	5.6	13.9
	Total*	80.4	19.6	100.0
<b>Writing</b>				
W3.1/3.2	Write compositions	0.0	13.8	13.8
W3.3	Proofread writing	41.7	2.8	44.5
W3.4	Revise writing for organization	41.7	0.0	41.7
	Total	83.4	16.6	100.0

\* The mathematics total does not include the B/C/D objective.

**Table 5 – Item Breakout by Content Performance Standard – HSGQE**

Content Area/Standard		Percent of Items		
Content Standard	Title	Multiple Choice	Constructed Response	Total
<b>Reading</b>				
R4.1	Use context clues	8.0	0.0	8.0
R4.4	Summarize information	24.0	4.0	28.0
R4.5	Critique arguments	18.0	6.0	24.0
R4.6	Apply multi-step directions	6.0	6.0	12.0
R4.9	Make and support assertions	6.0	8.0	14.0
R4.10	Analyze and evaluate themes	10.0	4.0	14.0
	Total	72.0	28.0	100.0
<b>Mathematics</b>				
B/C/D	Prob. Solve/Comm/Reasoning	0.0	6.9	6.9
M.A.1	Numeration	20.7	1.7	22.4
M.A.2	Measurement	17.2	3.5	20.7
M.A.3	Estimation & Computation	17.2	3.5	20.7
M.A.4	Functions & Relationships	10.3	1.7	12.0
M.A.5	Geometry	8.7	1.7	10.4
M.A.6	Statistics/Probability	12.1	1.7	13.8
	Total*	86.2	13.8	100.0
<b>Writing</b>				
W4.1/4.2	Write compositions	0.0	14.3	14.3
W4.3	Use conventional English	40.0	5.7	45.7
W4.4	Revise writing for word choice	34.3	5.7	40.0
	Total	74.3	25.7	100.0

\* The mathematics total does not include the B/C/D objective, which shares with other standards.

## Test Content Alignment

**Reading:** The Spring assessment 2000, Fall Retest assessment 2000, Spring assessment 2001, and the Fall Retest assessment 2001 (Forms A through D) were developed based on the original blueprint. This blueprint was revised during the summer of 2001 when the test was refocused to measure essential skills. The new blueprint eliminated two performance standards and made slight adjustments to the proportion of multiple-choice and constructed-response items in each of the remaining performance standards. The test length was shortened slightly. The new blueprint went into effect with the Spring assessment in 2002 (Form E).

**Math:** The Fall Retest assessment 2001 (Form D) was the last form that tested all of the content standards through the 15-18 year age range. The Spring 2002 assessment (Form E) was the first form to use the refocused content standards that measured essential skills established by Alaska. This eliminated some of the more advanced mathematical concepts. Form E did not hit the desired distribution across the standards because of our limited item pool. New items were field tested in Form E to enhance the item pool allowing the later refined forms to match the desired distribution.

**Writing:** The Spring assessment 2000, Fall Retest assessment 2000, and the Spring assessment 2001 (Forms A through C) were developed based on the original blueprint, while the Fall Retest in 2001 (Form D) was a repeat of Form A. This blueprint was revised during the summer of 2001 when the test was refocused to measure essential skills. The new blueprint called for the addition of new constructed-response item formats in one performance standard and made adjustments to the proportion of items across performance standards. The scores for items in the first performance standard are now double-weighted, thus achieving the desired content proportionality through two means while shortening test length. The new blueprint has been phased in through Spring and Fall 2002 (Forms E and F), as it was necessary to gather field test data on the new item formats; the Spring 2003 assessment (Form G) is the first form that matches the new blueprint exactly.

Tables 6-8 show content proportionality of HSGQE Form E for Reading, Mathematics, and Writing, respectively.

**Table 6 – Content Proportionality – HSGQE Spring 2003 Reading**

CONTENT	Test Items MC/SCR/ECR	Total Test Items	Test Points MC/SCR/ECR	Total Test Points Fraction (%)
R4.1: Use context clues	4/0/0	4	4/0/0	4/71 (6%)
R4.4: Summarize information	12/2/0	14	12/5/0	17/71 (24%)
R4.5: Critique arguments	9/3/0	12	9/6/0	15/71 (21%)
R4.6: Apply multi-step directions	3/2/1	6	3/5/4	12/71 (17%)
R4.9: Make and support assertions	3/4/0	7	3/10/0	13/71 (18%)
R4.10: Analyze and evaluate themes	5/2/0	7	5/5/0	10/71 (14%)
<b>TOTAL</b>	<b>36/13/1</b>	<b>50</b>	<b>36/31/4</b>	<b>71/71 (100%)</b>

**Table 7 – Content Proportionality – HSGQE Spring 2003 Mathematics**

<b>CONTENT</b>	<b>Test Items MC/SCR/ECR</b>	<b>Total Test Items</b>	<b>Test Points MC/SCR/ECR</b>	<b>Total Test Points fraction (%)</b>
A1: Numeration	12/1/0	13	12/2/0	14/70 (20%)
A2: Measurement	10/2/0	12	10/4/0	14/70 (20%)
A3: Estimation and Computation	10/2/0	12	10/4/0	14/70 (20%)
A4: Functions and Relationships	6/0/1	7	6/0/4	10/70 (14%)
A5: Geometry	5/1/0	6	5/2/0	7/70 (10%)
A6: Statistics/ Probability	7/0/1	8	7/0/4	11/70 (16%)
<b>TOTAL</b>	<b>50/6/2</b>	<b>58</b>	<b>50/12/8</b>	<b>70/70 (100.0%)</b>

**Table 8 – Content Proportionality – HSGQE Spring 2003 Writing**

<b>CONTENT</b>	<b>Test Items MC/SCR/ECR</b>	<b>Total Test Items</b>	<b>Test Points MC/SCR/ECR</b>	<b>Total Test Points fraction (%)</b>
W4.1/4.2: Write compositions	0/0/5	5	0/0/22	22/59 (37%)
W4.3: Use conventional English	14/0/2	16	14/0/8	22/59 (37%)
W4.4: Revise writing for word choice	12/2/0	14	12/3/0	15/59 (25%)
<b>TOTAL</b>	<b>26/2/7</b>	<b>35</b>	<b>26/3/30</b>	<b>59/59 (100%)</b>

## Item Selection

Item selection for the HSGQE and the Benchmark Assessments was completed by content editors and reviewed by research staff. The primary criterion for the selection of items was to meet the content specifications and statistical research guidelines. Within the limits set by these requirements, editors selected items with the best statistical characteristics. The criterion for minimizing measurement error throughout the expected range of performance results in items being chosen with a range of difficulties appropriate to the target grade. Such a procedure helps ensure that the resulting test does not exhibit floor or ceiling effects when examinees have either very low raw scores or perfect raw scores on the test (Lord, 1980).

### Item Selection Program

CTB editors and research staff use ItemSys (a microcomputer-based program) to select forms that (1) satisfy content requirements, (2) minimize bias levels, and (3) maximize psychometric integrity (i.e., reliability and range of achievement measured) within and across forms. The ItemSys program (Burket, 1988) creates an interactive connection between the developer selecting the test and the item database. The program monitors the impact of each decision made during the item selection process and offers the developer a variety of options for grouping, classifying, sorting, and ranking items to highlight key information as it is needed (Green, Yen, & Burket, 1989). The primary advantage of this computerized system is that it allows content editors to have complete knowledge and flexibility in selecting items which produce optimum test form characteristics, especially test characteristic and standard error of measurement functions. Content editors are able to do this while attending to the content requirements of parallel forms.

Item characteristics used by the content editors and maintained in the ItemSys database include

- ◆ the IRT parameters;
- ◆ the maximum information provided by each item;
- ◆ the estimated proportion correct by ability level;
- ◆ the contribution of each item to the accuracy of the test at every ability level;
- ◆ the floors and ceilings of each set of possible items that might be selected; and
- ◆ the estimated test form characteristics of any set of items selected.

In addition, ItemSys maintains flags for items which exhibit poor fit to the IRT model used as well as combined gender and ethnic differential item functioning indicators.

### Item Selection Process

The ItemSys program has three parts. The first part is used to select a working item pool of manageable size from the larger field test pool. Information on each item in the pool includes

- ◆ the content objective to which the item is assigned;
- ◆ a descriptive phrase about the item;
- ◆ the association of the item with a passage or stimulus;
- ◆ a bias rating indicating whether the item shows DIF to a particular population of students;



- ◆ parameters; and
- ◆ a fit rating indicating how well the item fits the expectations based on the IRT model used.

The second part of the program operates on this working pool to perform the actual test selection. Typically, the developer begins by specifying the number of items to be included in the test and a target number of items for each content objective. The computer is then prompted to automatically select a test that represents the best possible statistical combination of items. These automatic selections can then be used as a reference set to which other selections are compared. Successive selections are plotted together on a graphic display that shows the standard error of measurement for the two sets of selected items across the target grade range. This standard error of measurement curve becomes the visual cue the developer uses to adjust the selection. While identifying the ideal content, the developer simultaneously attempts to affect the curve positively. Lowering the curve translates into a test with less measurement error. Smoothing bumps out of the curve creates a test that measures more consistently across the target range of difficulty.

Moreover, the developer can at any time call up one of many information screens that rank items according to criteria such as their difficulty, their ability to discriminate between low- and high-achieving students, and their contribution to lowering the standard error curve at one particular point in the range. Such screens help the developer pinpoint the exact item to add to the selection or indicate when an optimum choice has already been made.

The third part of the program provides a table that shows both the expected number correct and the standard error of measurement as functions of scale score, as well as statistical and graphic summaries on bias, fit, and the average standard error of the test as finally selected. Any fault in the final selection becomes immediately apparent as the final statistics are generated. For instance, the developer can see whether the test is too easy or too difficult for the target grade, contains biased items, does not meet the requirements to match a parallel form, or does not adequately cover part of the range. If the developer detects any such problems, he or she can then return to the second stage of the program and revise the selection. The flexibility and utility of the program encourages multiple attempts at fine-tuning the selection.

A complete description of ItemSys, processes, and results to produce the operational forms can be found in the *Alaska Comprehensive System of Student Assessment: 1999 Technical Report for Benchmark Assessments and the High School Qualifying Exam*.

### Item Selection Results

The construction of the Alaska forms to be administered in 2003 required the fulfillment of content category quotas as well as the specified statistical/psychometric requirements. The resulting test forms selected exhibited the primary criterion of content. Within the limits set by these content requirements, review of the test selection determined that editors had selected items with the best statistical characteristics and minimal bias. Additionally, it was determined that the measurement error was minimized throughout the expected range of performance results by the selection of items with an appropriate range of difficulties.

Table 9 below provides a summary of the number of items with a specific difficulty. An item with difficulty between 0.00 and 0.09 would be considered extremely difficult, while an item with difficulty between 0.90 and 0.99 would be fairly easy. As the table shows, the majority of items were moderately difficult with difficulties ranging from 0.40 to 0.89.

**Table 9 – Item Difficulty Frequency Distribution (HSGQE)**





Difficulty Scale		Content Area		
Difficult	p-value	RD	WR	MA
	0.00 – 0.09	0	0	0
	0.10 – 0.19	1	0	0
	0.20 – 0.29	0	0	2
	0.30 – 0.39	2	0	2
	0.40 – 0.49	4	1	10
	0.50 – 0.59	4	3	11
	0.60 – 0.69	9	7	12
	0.70 – 0.79	14	10	12
	0.80 – 0.89	12	12	7
	0.90 – 0.99	4	2	2
	Total	50	35	58
Easy				

Table 10 shows frequency distributions for item difficulty by item type for each content area. This table indicates that the majority of the test items are in the easy to medium difficulty range, with only a few items categorized as difficult.

**Table 10 – Item Difficulty Frequency Distribution by Score Points for HSGQE Spring 2003**

Content Area		Item Range	Multiple Choice and Constructed Response Item Score Points							
Difficult to Easy			MC	1 pt.	2 pts.	3 pts.	4 pts.	5 pts.	6 pts.	Total
RD		0.00 – 0.09								
		0.10 – 0.19			1					1
		0.20 – 0.29								
		0.30 – 0.39	1			1				2
		0.40 – 0.49	2		1		1			4
		0.50 – 0.59	2		1	1				4
		0.60 – 0.69	5		2	2				9
		0.70 – 0.79	10		3	1				14
		0.80 – 0.89	12							12
		0.90– 0.99	4							4
Total			36		8	5	1			50
WR		0.00 – 0.09								
		0.10 – 0.19								
		0.20 – 0.29								
		0.30 – 0.39								
		0.40 – 0.49			1					1
		0.50 – 0.59	1				1		1	3
		0.60 – 0.69	1	1			5			7
		0.70 – 0.79	10							10
		0.80 – 0.89	12							12
		0.90– 0.99	2							2
Total			26	1	1		6		1	35
MA		0.00 – 0.09								
		0.10 – 0.19								
		0.20 – 0.29			2					2
		0.30 – 0.39	2							2
		0.40 – 0.49	8		2					10
		0.50 – 0.59	9		1		1			11
		0.60 – 0.69	11				1			12
		0.70 – 0.79	11		1					12
		0.80 – 0.89	7							7
		0.90– 0.99	2							2
Total			50		6		2			58

### Individual Item Analyses

Tables 11– 22 present individual item data for each content area and grade level tested in the spring of 2003. The tables include P-Values for each item as well as Point Biserial information on each possible distracter. The percentage of the total testing population choosing a particular answer option for a particular item is also listed. This information is provided for all items. The constructed-response items are listed only with their P-Values.

Table 11 – Benchmark 1 Reading Item Statistics

Benchmark 1 Grade 3 Reading Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biseal for Each Option			
		A	B	C	D	A	B	C	D
1	0.950	*94.98%	2.03%	2.84%		0.395	-0.269	-0.335	
2	0.947	1.20%	*94.67%	3.92%		-0.246	0.286	-0.232	
3	0.759	*75.94%	4.91%	18.72%		0.476	-0.377	-0.405	
4	0.853	4.16%	*85.32%	10.38%		-0.285	0.424	-0.390	
5	0.879	6.84%	5.06%	*87.90%		-0.262	-0.447	0.441	
6	0.905	5.63%	*90.47%	3.65%		-0.354	0.479	-0.388	
7	0.871	*87.12%	8.59%	3.98%		0.423	-0.324	-0.365	
8	0.837	9.68%	4.39%	*83.65%	2.05%	-0.451	-0.275	0.558	-0.291
9	0.899	4.33%	1.49%	3.93%	*89.92%	-0.355	-0.260	-0.270	0.466
10	0.650	17.90%	11.22%	*64.96%	5.47%	-0.178	-0.306	0.399	-0.340
11	0.912	3.26%	2.39%	2.36%	*91.23%	-0.283	-0.239	-0.290	0.423
12	0.327	37.41%	*32.75%	18.74%	9.91%	-0.160	0.163	-0.106	-0.122
13	0.697	*69.71%	5.51%	8.15%	14.44%	0.446	-0.276	-0.268	-0.302
14	0.766	10.53%	*76.56%	4.05%	5.35%	-0.387	0.550	-0.299	-0.351
15	0.635								
16	0.490								
17	0.797								
18	0.527	*52.67%	28.52%	12.11%	4.91%	0.204	-0.089	-0.169	-0.262
19	0.728	9.94%	12.89%	3.42%	*72.81%	-0.145	-0.291	-0.332	0.381
20	0.586	10.59%	13.65%	15.71%	*58.65%	-0.329	-0.191	-0.227	0.415
21	0.676	*67.61%	3.64%	4.43%	21.24%	0.499	-0.330	-0.323	-0.302
22	0.383								
23	0.879	7.33%	*87.92%	3.54%		-0.438	0.541	-0.326	
24	0.878	5.03%	*87.78%	4.33%		-0.362	0.524	-0.411	
25	0.778	*77.77%	7.36%	13.86%		0.530	-0.426	-0.360	
26	0.828	*82.84%	10.36%	5.74%		0.564	-0.412	-0.402	
27	0.834	8.93%	6.51%	*83.37%		-0.357	-0.336	0.472	
28	0.808	6.95%	8.16%	*80.85%	2.48%	-0.285	-0.430	0.595	-0.294
29	0.871	4.07%	2.39%	*87.13%	4.68%	-0.288	-0.133	0.419	-0.276
30	0.761	7.49%	*76.08%	10.15%	5.10%	-0.289	0.548	-0.344	-0.319
31	0.830	*83.03%	6.67%	5.40%	3.48%	0.550	-0.297	-0.367	-0.284
32	0.565	4.87%	3.02%	*56.52%	34.02%	-0.321	-0.315	0.264	-0.101
33	0.294								
34	0.583	*58.31%	6.62%	22.65%	10.91%	0.358	-0.302	-0.146	-0.259
35	0.619	6.24%	21.44%	8.75%	*61.89%	-0.241	-0.224	-0.239	0.377
36	0.279								

\* indicates the correct answer

Table 12 – Benchmark 1 Mathematics Item Statistics

Benchmark 1 Grade 3 Mathematics Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserial for Each Option			
		A	B	C	D	A	B	C	D
1	0.920	3.29%	2.73%	*92.03%	1.37%	-0.274	-0.202	0.346	-0.175
2	0.748	8.55%	7.28%	*74.78%	8.60%	-0.263	-0.252	0.408	-0.274
3	0.776	*77.62%	12.41%	5.57%	3.74%	0.405	-0.297	-0.256	-0.233
4	0.866	6.33%	2.51%	3.78%	*86.60%	-0.330	-0.208	-0.308	0.451
5	0.542	22.94%	11.70%	10.70%	*54.17%	-0.264	-0.236	-0.127	0.318
6	0.891								
7	0.783								
8	0.744	1.63%	*74.44%	13.85%	9.34%	-0.147	0.349	-0.272	-0.262
9	0.850	6.78%	2.70%	4.60%	*84.99%	-0.213	-0.198	-0.216	0.312
10	0.869	2.44%	8.15%	*86.92%	1.45%	-0.203	-0.251	0.315	-0.173
11	0.620	*62.02%	13.68%	6.73%	16.69%	0.437	-0.177	-0.224	-0.389
12	0.636	4.05%	*63.56%	21.50%	10.27%	-0.272	0.412	-0.197	-0.371
13	0.588	6.25%	18.84%	15.54%	*58.81%	-0.178	-0.328	-0.334	0.480
14	0.741	13.88%	7.63%	*74.14%	3.66%	-0.307	-0.288	0.437	-0.240
15	0.611	21.85%	7.30%	8.64%	*61.09%	-0.190	-0.252	-0.265	0.342
16	0.591	6.42%	12.67%	*59.14%	21.01%	-0.218	-0.245	0.378	-0.265
17	0.752								
18	0.829								
19	0.810	8.12%	*80.99%	3.91%	6.06%	-0.197	0.403	-0.226	-0.352
20	0.639	*63.86%	17.64%	5.87%	11.52%	0.393	-0.179	-0.207	-0.367
21	0.800	*79.95%	5.61%	6.86%	4.27%	0.468	-0.263	-0.302	-0.268
22	0.863	7.39%	3.42%	*86.25%	2.25%	-0.216	-0.193	0.313	-0.230
23	0.534	5.63%	*53.36%	18.83%	21.38%	-0.210	0.500	-0.207	-0.423
24	0.720	*71.95%	5.81%	4.27%	17.02%	0.439	-0.254	-0.221	-0.346
25	0.729	20.74%	4.36%	*72.91%	1.17%	-0.450	-0.121	0.406	-0.086
26	0.710	2.52%	5.22%	20.66%	*71.03%	-0.153	-0.288	-0.273	0.346
27	0.554	6.01%	19.78%	17.88%	*55.39%	-0.312	-0.277	-0.147	0.374
28	0.375	23.42%	26.29%	*37.53%	12.05%	-0.280	-0.040	0.299	-0.253
29	0.333								
30	0.403								
31	0.762	8.35%	6.56%	*76.21%	8.17%	-0.422	-0.207	0.496	-0.268
32	0.607	*60.67%	15.08%	11.42%	10.84%	0.430	-0.293	-0.230	-0.267
33	0.787	11.27%	5.61%	*78.71%	3.63%	-0.374	-0.273	0.501	-0.260
34	0.665	17.10%	10.70%	4.82%	*66.48%	-0.441	-0.232	-0.112	0.465
35	0.641	7.19%	*64.12%	10.73%	17.15%	-0.262	0.411	-0.236	-0.276
36	0.767	4.88%	9.33%	7.34%	*76.70%	-0.297	-0.280	-0.293	0.467

\* indicates the correct answer

Table 13 – Benchmark 1 Writing Item Statistics

Benchmark 1 Grade 3 Writing Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserial for Each Option			
		A	B	C	D	A	B	C	D
1	0.539	38.08%	7.69%	*53.91%		-0.385	-0.316	0.456	
2	0.726	*72.60%	7.85%	13.53%	5.38%	0.478	-0.252	-0.338	-0.273
3	0.712	17.32%	10.68%	*71.17%		-0.352	-0.335	0.456	
4	0.440	28.16%	*43.99%	7.93%	17.24%	-0.173	0.339	-0.187	-0.238
5	0.635	30.84%	4.84%	*63.50%		-0.254	-0.298	0.290	
6	0.691	11.69%	6.09%	11.95%	*69.06%	-0.324	-0.245	-0.280	0.474
7	0.640	1.63%	32.40%	1.70%	*63.95%	-0.195	-0.250	-0.217	0.265
8	0.866	5.04%	1.79%	*86.56%	5.97%	-0.327	-0.225	0.486	-0.330
9	0.466								
10	0.575	26.56%	9.42%	5.94%	*57.51%	-0.218	-0.287	-0.264	0.396
11	0.730	*73.04%	7.24%	4.17%	14.75%	0.394	-0.210	-0.249	-0.288
12	0.553								
13	0.543	*54.33%	22.64%	15.59%	6.71%	0.348	-0.182	-0.191	-0.309
14	0.363	22.36%	26.18%	14.06%	*36.30%	-0.136	-0.132	-0.221	0.276
15	0.665	*66.54%	6.15%	17.02%	8.97%	0.454	-0.271	-0.266	-0.281
16	0.569	23.50%	*56.90%	15.70%	3.08%	-0.250	0.285	-0.117	-0.232
17	0.601								
18	0.593	3.39%	*59.26%	34.66%	2.28%	-0.271	0.432	-0.349	-0.231
19	0.814	14.25%	3.80%	*81.44%		-0.397	-0.261	0.439	
20	0.893	2.67%	3.60%	*89.26%	3.85%	-0.233	-0.231	0.430	-0.314
21	0.912	3.35%	*91.22%	1.60%	2.83%	-0.243	0.374	-0.178	-0.252
22	0.781	*78.09%	6.67%	5.35%	9.35%	0.397	-0.272	-0.208	-0.250
23	0.449	35.26%	7.89%	11.26%	*44.86%	-0.110	-0.281	-0.277	0.331
24	0.346	39.90%	*34.63%	14.87%	9.75%	-0.051	0.221	-0.181	-0.241
25	0.373								
26	0.237								
27	0.806	7.93%	6.21%	*80.56%	4.47%	-0.277	-0.303	0.475	-0.265
28	0.482	28.13%	7.09%	*48.23%	15.57%	-0.136	-0.295	0.233	-0.076
29	0.653	*65.29%	23.53%	4.00%	6.26%	0.469	-0.365	-0.289	-0.163
30	0.935	*93.48%	1.47%	1.99%	1.99%	0.384	-0.197	-0.209	-0.261
31	0.849	*84.90%	5.47%	4.08%	3.95%	0.520	-0.306	-0.325	-0.290
32	0.556								
33	0.508	10.22%	28.29%	*50.80%	9.50%	-0.336	-0.276	0.454	-0.146
34	0.729	7.22%	*72.91%	14.47%	4.03%	-0.254	0.408	-0.235	-0.282
35	0.400	6.15%	32.74%	*40.02%	19.86%	-0.251	-0.205	0.343	-0.152
36	0.380	16.21%	9.23%	*37.98%	34.73%	-0.214	-0.247	0.358	-0.151

\* indicates the correct answer

Table 14 – Benchmark 2 Reading Item Statistics

Benchmark 2 Grade 6 Reading Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biseriars for Each Option			
		A	B	C	D	A	B	C	D
1	0.493	32.37%	13.87%	*49.3%	4.38%	-0.080	-0.334	0.241	-0.239
2	0.850	*85.00%	7.41%	3.54%	3.81%	0.373	-0.280	-0.257	-0.226
3	0.839	*83.86%	4.01%	1.63%	10.41%	0.372	-0.278	-0.232	-0.282
4	0.873	3.63%	*87.29%	6.25%	2.62%	-0.242	0.436	-0.363	-0.227
5	0.802	2.25%	*80.18%	4.29%	12.97%	-0.237	0.448	-0.303	-0.350
6	0.958	1.10%	0.85%	2.12%	*95.75%	-0.173	-0.221	-0.283	0.354
7	0.804	5.49%	*80.42%	9.23%	4.60%	-0.228	0.398	-0.234	-0.370
8	0.758	7.25%	*75.83%	4.98%	11.79%	-0.276	0.382	-0.334	-0.206
9	0.586								
10	0.917	2.50%	1.29%	*91.67%	4.22%	-0.288	-0.253	0.455	-0.329
11	0.832	6.49%	*83.21%	5.38%	4.69%	-0.304	0.412	-0.291	-0.217
12	0.912	4.19%	3.90%	*91.22%	0.50%	-0.259	-0.273	0.339	-0.124
13	0.869	5.62%	1.95%	*86.90%	5.31%	-0.316	-0.244	0.457	-0.320
14	0.144								
15	0.762	2.90%	10.86%	*76.21%	9.09%	-0.248	-0.330	0.449	-0.289
16	0.930	2.99%	0.56%	*93.00%	2.99%	-0.324	-0.125	0.367	-0.234
17	0.651	6.16%	*65.09%	6.55%	21.92%	-0.357	0.520	-0.250	-0.373
18	0.735	0.73%	5.59%	19.82%	*73.55%	-0.199	-0.320	-0.436	0.503
19	0.597	19.32%	*59.69%	11.06%	9.74%	-0.369	0.368	-0.256	-0.078
20	0.838	*83.80%	7.49%	5.58%	2.75%	0.518	-0.341	-0.374	-0.256
21	0.696	6.40%	13.94%	*69.63%	9.48%	-0.211	-0.194	0.300	-0.263
22	0.847	7.11%	*84.68%	4.84%	2.71%	-0.279	0.429	-0.298	-0.265
23	0.859	3.91%	4.13%	5.21%	*85.93%	-0.230	-0.306	-0.314	0.451
24	0.566	*56.63%	16.88%	11.70%	13.99%	0.398	-0.218	-0.273	-0.275
25	0.261								
26	0.664	*66.42%	10.45%	15.95%	6.89%	0.303	-0.158	-0.254	-0.253
27	0.731	14.53%	6.00%	*73.07%	6.09%	-0.378	-0.331	0.558	-0.327
28	0.866	2.55%	2.16%	8.34%	*86.63%	-0.300	-0.261	-0.423	0.532
29	0.881	*88.12%	4.52%	2.78%	4.10%	0.469	-0.281	-0.295	-0.321
30	0.522	18.37%	14.14%	*52.20%	14.96%	-0.230	-0.286	0.386	-0.227
31	0.401	15.88%	*40.09%	8.30%	35.09%	-0.248	0.279	-0.275	-0.111
32	0.794	*79.38%	7.73%	4.27%	8.09%	0.557	-0.374	-0.314	-0.347
33	0.524	13.36%	24.36%	*52.37%	8.80%	-0.187	-0.271	0.367	-0.264
34	0.426								
35	0.224								
36	0.369								

\* indicates the correct answer

Table 15 – Benchmark 2 Mathematics Item Statistics

Benchmark 2 Grade 6 Mathematics Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserial for Each Option			
		A	B	C	D	A	B	C	D
1	0.881	2.64%	*88.13%	2.09%	6.83%	-0.180	0.179	-0.062	-0.155
2	0.899	1.52%	4.05%	4.33%	*89.92%	-0.140	-0.208	-0.220	0.287
3	0.796	8.32%	*79.61%	8.01%	3.01%	-0.374	0.470	-0.295	-0.161
4	0.889	2.03%	2.42%	*88.92%	6.08%	-0.108	-0.184	0.243	-0.201
5	0.874	*87.38%	1.76%	5.48%	3.59%	0.405	-0.167	-0.336	-0.249
6	0.866	3.00%	6.66%	*86.60%	2.18%	-0.215	-0.301	0.350	-0.149
7	0.908	1.94%	3.33%	*90.76%	3.80%	-0.144	-0.262	0.375	-0.290
8	0.804	6.96%	*80.40%	5.71%	6.76%	-0.269	0.370	-0.311	-0.141
9	0.888	0.91%	3.02%	6.94%	*88.84%	-0.157	-0.165	-0.324	0.349
10	0.584								
11	0.555								
12	0.768	3.39%	4.63%	*76.83%	14.93%	-0.239	-0.219	0.475	-0.396
13	0.820	10.14%	3.22%	4.33%	*81.96%	-0.294	-0.226	-0.232	0.394
14	0.510	12.99%	22.56%	12.74%	*51.02%	-0.192	-0.339	-0.239	0.466
15	0.599	4.90%	27.82%	*59.95%	6.91%	-0.109	-0.511	0.528	-0.182
16	0.672	4.92%	*67.18%	22.24%	5.38%	-0.189	0.348	-0.303	-0.175
17	0.806	*80.62%	9.42%	7.43%	2.32%	0.496	-0.299	-0.413	-0.175
18	0.553	23.96%	6.30%	14.15%	*55.31%	-0.155	-0.305	-0.248	0.346
19	0.231								
20	0.525								
21	0.737	4.24%	6.90%	14.77%	*73.66%	-0.300	-0.278	-0.310	0.475
22	0.801	*80.10%	6.44%	10.84%	1.98%	0.480	-0.283	-0.374	-0.200
23	0.752	5.50%	6.02%	*75.22%	12.81%	-0.172	-0.319	0.402	-0.272
24	0.683	*68.28%	23.68%	6.10%	1.51%	0.399	-0.269	-0.369	-0.167
25	0.777	9.03%	*77.70%	4.69%	7.72%	-0.333	0.476	-0.238	-0.297
26	0.720	*72.05%	10.35%	10.69%	4.61%	0.431	-0.299	-0.289	-0.202
27	0.457	2.39%	4.45%	45.51%	*45.72%	-0.208	-0.202	-0.407	0.478
28	0.462								
29	0.155								
30	0.620	*62.01%	12.94%	15.34%	9.28%	0.474	-0.170	-0.282	-0.396
31	0.489	*48.89%	10.81%	34.58%	5.31%	0.527	-0.206	-0.438	-0.173
32	0.566								
33	0.326								
34	0.760	2.54%	*76.02%	15.75%	4.67%	-0.225	0.522	-0.387	-0.334
35	0.602	4.34%	*60.23%	2.75%	32.44%	-0.245	0.389	-0.224	-0.317
36	0.625	10.76%	*62.51%	12.18%	13.87%	-0.341	0.346	-0.177	-0.149

\* indicates the correct answer



Table 16 – Benchmark 2 Writing Item Statistics

Benchmark 2 Grade 6 Writing Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserials for Each Option			
		A	B	C	D	A	B	C	D
1	0.741	*74.06%	11.01%	11.08%	3.74%	0.493	-0.296	-0.357	-0.242
2	0.753	21.84%	1.53%	*75.31%	1.24%	-0.229	-0.202	0.245	-0.200
3	0.603	28.22%	*60.28%	4.55%	6.86%	-0.277	0.380	-0.237	-0.243
4	0.461	7.55%	21.87%	*46.12%	23.80%	-0.261	-0.275	0.341	-0.116
5	0.543								
6	0.778	3.12%	7.54%	*77.79%	11.21%	-0.143	-0.287	0.373	-0.270
7	0.887	4.10%	*88.66%	3.98%	2.93%	-0.244	0.397	-0.256	-0.248
8	0.294	*29.45%	6.89%	48.26%	14.82%	0.301	-0.290	-0.196	-0.065
9	0.608								
10	0.538	19.18%	21.76%	4.89%	*53.82%	-0.210	-0.296	-0.238	0.409
11	0.688	1.87%	*68.79%	7.29%	21.67%	-0.204	0.211	-0.260	-0.110
12	0.795	*79.47%	10.63%	6.68%	2.28%	0.333	-0.241	-0.227	-0.206
13	0.562								
14	0.751	9.95%	10.07%	*75.09%	4.49%	-0.187	-0.401	0.439	-0.219
15	0.673	10.98%	16.10%	*67.27%	4.36%	-0.279	-0.224	0.371	-0.226
16	0.765	*76.52%	12.79%	3.47%	6.92%	0.462	-0.326	-0.293	-0.250
17	0.798	7.57%	9.10%	*79.81%	3.19%	-0.242	-0.278	0.395	-0.238
18	0.742	3.20%	3.50%	18.36%	*74.19%	-0.281	-0.281	-0.061	0.208
19	0.488								
20	0.744	*74.43%	12.55%	6.49%	6.17%	0.447	-0.323	-0.280	-0.207
21	0.601	*60.05%	11.77%	20.65%	5.77%	0.397	-0.315	-0.175	-0.279
22	0.669	9.88%	6.87%	*66.91%	15.92%	-0.262	-0.327	0.430	-0.219
23	0.682	6.44%	16.54%	8.06%	*68.22%	-0.248	-0.260	-0.300	0.433
24	0.435	11.79%	9.34%	*43.48%	34.97%	-0.189	-0.234	0.296	-0.158
25	0.702	*70.18%	9.43%	15.28%	4.61%	0.454	-0.336	-0.268	-0.222
26	0.861	8.55%	3.23%	*86.14%	1.69%	-0.203	-0.292	0.354	-0.230
27	0.831	*83.12%	4.62%	5.90%	5.67%	0.507	-0.309	-0.335	-0.269
28	0.851	2.82%	4.04%	*85.12%	5.22%	-0.267	-0.323	0.443	-0.284
29	0.598								
30	0.521	15.23%	28.71%	*52.12%	3.29%	-0.229	-0.275	0.391	-0.203
31	0.633	8.28%	5.66%	21.47%	*63.29%	-0.198	-0.304	-0.254	0.396
32	0.619	12.06%	7.20%	*61.92%	18.28%	-0.279	-0.363	0.372	-0.100
33	0.828	2.48%	*82.85%	7.02%	7.09%	-0.257	0.423	-0.277	-0.259
34	0.833	3.68%	7.58%	*83.30%	4.90%	-0.283	-0.346	0.495	-0.257
35	0.504								
36	0.550								

\* indicates the correct answer

Table 17 – Benchmark 3 Reading Item Statistics

Benchmark 3 Grade 8 Reading Item Statistics									
ITEM	P-Value	Percentage of Total Selecting Each Option				Point Biserials for Each Option			
		A	B	C	D	A	B	C	D
1	0.872	1.33%	*87.21%	4.69%	6.66%	-0.225	0.437	-0.193	-0.415
2	0.941	2.40%	2.05%	*94.14%	1.29%	-0.195	-0.267	0.356	-0.242
3	0.821	10.58%	1.89%	*82.08%	5.32%	-0.323	-0.203	0.325	-0.146
4	0.796	1.29%	10.01%	*79.62%	8.87%	-0.235	-0.347	0.395	-0.229
5	0.630	*62.98%	20.20%	12.62%	4.11%	0.321	-0.247	-0.255	-0.175
6	0.932	*93.19%	2.55%	2.48%	1.68%	0.348	-0.253	-0.245	-0.179
7	0.925	2.37%	*92.53%	1.44%	3.52%	-0.224	0.394	-0.234	-0.300
8	0.830	8.45%	*82.99%	5.28%	3.10%	-0.313	0.487	-0.358	-0.262
9	0.816	4.32%	*81.61%	6.65%	7.22%	-0.343	0.505	-0.277	-0.339
10	0.475								
11	0.657	*65.67%	9.44%	18.57%	6.09%	0.453	-0.428	-0.233	-0.223
12	0.423								
13	0.250								
14	0.479	11.40%	18.80%	*47.90%	20.75%	-0.215	-0.194	0.377	-0.274
15	0.909	2.71%	*90.93%	3.72%	2.41%	-0.225	0.434	-0.304	-0.290
16	0.693	*69.35%	11.32%	13.61%	5.38%	0.444	-0.255	-0.345	-0.237
17	0.663	3.26%	19.50%	*66.25%	10.53%	-0.119	-0.233	0.324	-0.308
18	0.644	5.94%	9.77%	*64.36%	18.98%	-0.269	-0.221	0.400	-0.298
19	0.920	2.35%	4.56%	*92.04%	0.68%	-0.155	-0.209	0.256	-0.152
20	0.849	1.54%	8.95%	4.09%	*84.94%	-0.224	-0.424	-0.305	0.535
21	0.868	1.25%	7.15%	*86.76%	4.30%	-0.217	-0.243	0.378	-0.299
22	0.720	4.45%	3.32%	19.71%	*71.99%	-0.226	-0.242	-0.325	0.398
23	0.908	*90.79%	1.92%	2.00%	4.93%	0.394	-0.215	-0.261	-0.276
24	0.775	6.18%	*77.46%	2.39%	13.53%	-0.283	0.451	-0.239	-0.340
25	0.822	5.30%	6.41%	5.62%	*82.18%	-0.248	-0.301	-0.283	0.440
26	0.744								
27	0.849	7.99%	*84.93%	5.36%	1.27%	-0.374	0.502	-0.313	-0.230
28	0.603	5.96%	*60.29%	27.96%	5.12%	-0.290	0.437	-0.321	-0.236
29	0.584								
30	0.748	4.87%	*74.83%	11.71%	7.98%	-0.244	0.412	-0.284	-0.266
31	0.538	5.67%	24.93%	14.77%	*53.82%	-0.263	-0.190	-0.348	0.415
32	0.808	4.45%	*80.79%	10.59%	3.18%	-0.289	0.490	-0.362	-0.242
33	0.608								
34	0.737	15.33%	3.83%	*73.69%	6.69%	-0.333	-0.329	0.437	-0.187
35	0.703	10.85%	*70.30%	12.74%	5.54%	-0.218	0.422	-0.352	-0.223
36	0.833	5.90%	6.04%	*83.25%	4.13%	-0.257	-0.279	0.388	-0.206

\* indicates the correct answer

Table 18 – Benchmark 3 Mathematics Item Statistics

Benchmark 3 Grade 8 Mathematics Item Statistics									
ITEM	P-Value	Percentage of Total Selecting Each Option				Point Biserial For Each Option			
		A	B	C	D	A	B	C	D
1	0.947	*94.71%	0.98%	2.27%	0.61%	0.272	-0.145	-0.231	-0.106
2	0.855	4.05%	5.51%	4.78%	*85.48%	-0.321	-0.330	-0.266	0.491
3	0.759	9.25%	*75.93%	13.79%	0.57%	-0.323	0.475	-0.381	-0.120
4	0.845	3.70%	2.84%	8.73%	*84.50%	-0.171	-0.164	-0.263	0.299
5	0.806	1.15%	11.10%	6.91%	*80.55%	-0.179	-0.252	-0.395	0.428
6	0.714	9.59%	17.41%	*71.40%	1.20%	-0.332	-0.448	0.553	-0.164
7	0.608	12.13%	14.35%	*60.81%	12.29%	-0.333	-0.296	0.501	-0.251
8	0.804	*80.37%	5.58%	3.27%	9.65%	0.464	-0.266	-0.211	-0.366
9	0.777	16.53%	*77.69%	3.56%	1.28%	-0.420	0.458	-0.264	-0.094
10	0.729	8.23%	11.30%	*72.94%	6.59%	-0.258	-0.345	0.474	-0.255
11	0.435								
12	0.633	*63.35%	9.39%	24.54%	2.41%	0.425	-0.235	-0.367	-0.179
13	0.503	21.88%	*50.29%	17.58%	9.29%	-0.219	0.210	-0.249	0.050
14	0.619	6.29%	12.64%	*61.90%	18.66%	-0.119	-0.192	0.385	-0.372
15	0.598	12.25%	*59.81%	13.71%	13.30%	-0.280	0.420	-0.272	-0.227
16	0.585	10.07%	18.84%	*58.48%	11.82%	-0.143	-0.263	0.404	-0.331
17	0.785	5.81%	*78.52%	12.61%	2.50%	-0.197	0.354	-0.312	-0.170
18	0.572	11.29%	23.37%	*57.22%	6.35%	-0.294	-0.255	0.410	-0.216
19	0.678	8.23%	10.38%	11.72%	*67.76%	-0.265	-0.291	-0.349	0.511
20	0.707	13.52%	8.33%	4.59%	*70.74%	-0.332	-0.280	-0.253	0.479
21	0.261								
22	0.820								
23	0.288								
24	0.803	4.78%	6.16%	*80.29%	8.07%	-0.204	-0.335	0.499	-0.358
25	0.713	*71.29%	7.58%	16.32%	4.17%	0.531	-0.250	-0.425	-0.258
26	0.441	32.41%	13.90%	8.84%	*44.06%	-0.165	-0.282	-0.317	0.415
27	0.765	*76.45%	10.46%	9.02%	3.54%	0.506	-0.384	-0.308	-0.205
28	0.655	*65.53%	11.79%	18.20%	3.79%	0.536	-0.360	-0.337	-0.254
29	0.410	15.08%	*41.05%	31.20%	12.01%	-0.255	0.339	-0.227	-0.112
30	0.235								
31	0.855	4.56%	1.79%	7.72%	*85.55%	-0.265	-0.207	-0.268	0.380
32	0.567	13.39%	6.05%	*56.74%	23.33%	-0.318	-0.103	0.366	-0.246
33	0.146								
34	0.776	*77.62%	13.36%	3.56%	4.30%	0.405	-0.278	-0.190	-0.296
35	0.621	15.31%	11.85%	*62.07%	9.95%	-0.205	-0.354	0.473	-0.284
36	0.159								

\* indicates the correct answer

Table 19 – Benchmark 3 Writing Item Statistics

Benchmark 3 Grade 8 Writing Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserial for Each Option			
		A	B	C	D	A	B	C	D
1	0.948	*94.79%	1.63%	2.55%	0.96%	0.354	-0.234	-0.248	-0.184
2	0.798	6.94%	*79.78%	7.42%	5.42%	-0.243	0.382	-0.302	-0.177
3	0.797	3.97%	3.98%	*79.68%	11.58%	-0.226	-0.286	0.363	-0.226
4	0.791	3.24%	*79.11%	5.17%	12.40%	-0.242	0.312	-0.253	-0.184
5	0.600	*59.95%	3.76%	18.39%	17.49%	0.291	-0.269	-0.135	-0.242
6	0.494								
7	0.750	10.24%	*75.00%	6.04%	8.37%	-0.135	0.272	-0.228	-0.202
8	0.799	6.74%	10.03%	2.86%	*79.95%	-0.213	-0.337	-0.253	0.437
9	0.757	*75.75%	2.38%	3.55%	17.88%	0.345	-0.145	-0.254	-0.279
10	0.544								
11	0.746	8.57%	5.52%	*74.56%	9.63%	-0.215	-0.271	0.415	-0.239
12	0.612								
13	0.645	25.48%	3.24%	6.28%	*64.47%	-0.242	-0.318	-0.275	0.402
14	0.711	19.00%	6.48%	2.77%	*71.06%	-0.279	-0.302	-0.244	0.430
15	0.775	*77.47%	10.31%	3.82%	7.56%	0.467	-0.268	-0.276	-0.309
16	0.782	4.35%	3.56%	*78.23%	13.36%	-0.279	-0.303	0.332	-0.142
17	0.556	7.86%	*55.59%	7.77%	27.90%	-0.149	0.306	-0.196	-0.236
18	0.576								
19	0.546	29.10%	11.05%	4.76%	*54.57%	-0.165	-0.332	-0.203	0.355
20	0.570	2.64%	35.61%	*57.02%	3.95%	-0.267	-0.270	0.367	-0.244
21	0.640	13.65%	9.01%	*63.98%	12.86%	-0.168	-0.306	0.259	-0.061
22	0.692	8.46%	12.91%	8.85%	*69.22%	-0.231	-0.306	-0.300	0.481
23	0.822	*82.16%	6.57%	7.75%	3.00%	0.424	-0.283	-0.260	-0.226
24	0.632	10.65%	7.61%	*63.18%	17.90%	-0.221	-0.310	0.376	-0.183
25	0.528	5.13%	21.71%	*52.76%	19.77%	-0.220	-0.163	0.292	-0.199
26	0.422	15.18%	*42.24%	28.49%	13.39%	-0.140	0.292	-0.176	-0.206
27	0.648	13.51%	11.81%	*64.81%	8.62%	-0.221	-0.222	0.415	-0.312
28	0.769	*76.94%	7.63%	7.05%	6.45%	0.437	-0.308	-0.295	-0.148
29	0.583								
30	0.329	3.50%	27.04%	*32.90%	35.75%	-0.207	-0.097	0.158	-0.099
31	0.788	*78.77%	5.10%	12.16%	2.97%	0.398	-0.268	-0.236	-0.253
32	0.679	4.91%	*67.88%	20.72%	4.51%	-0.244	0.378	-0.233	-0.256
33	0.447	25.59%	*44.71%	7.01%	21.21%	-0.291	0.370	-0.226	-0.116
34	0.791	4.46%	11.65%	*79.07%	3.75%	-0.236	-0.340	0.429	-0.173
35	0.721	10.09%	5.28%	10.01%	*72.05%	-0.268	-0.290	-0.270	0.456
36	0.514								

\* indicates the correct answer

Table 20 – HSGQE Reading Item Statistics

HSGQE Reading Item Statistics									
ITEM	P-Value	Percent of Total Selecting for Each Option				Point Biserials For Each Option			
		A	B	C	D	A	B	C	D
1	0.86	6.60%	3.88%	3.09%	*85.892%	-0.39	-0.30	-0.26	0.54
2	0.68	15.21%	5.40%	*68.25%	10.83%	-0.29	-0.23	0.37	-0.14
3	0.94	1.49%	2.16%	1.92%	*94.28%	-0.26	-0.22	-0.23	0.39
4	0.18								
5	0.68								
6	0.57	21.74%	4.80%	*57.07%	15.58%	-0.16	-0.29	0.32	-0.18
7	0.87	*87.49%	3.17%	5.70%	2.02%	0.36	-0.21	-0.29	-0.17
8	0.82	4.26%	11.62%	2.34%	*81.51%	-0.27	-0.32	-0.28	0.48
9	0.43	41.86%	7.76%	*42.63%	7.31%	-0.18	-0.15	0.31	-0.25
10	0.63								
11	0.83	1.22%	*83.35%	10.74%	4.45%	-0.20	0.27	-0.19	-0.17
12	0.66	8.81%	15.29%	*66.08%	9.42%	-0.23	-0.21	0.33	-0.15
13	0.83	*82.98%	4.33%	6.00%	6.34%	0.49	-0.27	-0.29	-0.28
14	0.84	7.35%	5.59%	3.15%	*83.62%	-0.23	-0.28	-0.31	0.46
15	0.46	*46.46%	14.17%	21.30%	17.62%	0.40	-0.15	-0.31	-0.15
16	0.64								
17	0.77	9.14%	7.78%	*76.60%	5.30%	-0.20	-0.27	0.41	-0.23
18	0.49								
19	0.74	3.59%	12.12%	9.74%	*73.70%	-0.17	-0.24	-0.29	0.42
20	0.86	3.00%	*86.46%	4.05%	5.30%	-0.27	0.52	-0.29	-0.30
21	0.91	2.61%	*91.12%	1.39%	4.48%	-0.26	0.42	-0.24	-0.25
22	0.87	3.90%	3.18%	5.16%	*87.22%	-0.29	-0.27	-0.36	0.54
23	0.70								
24	0.74	*73.86%	20.04%	2.80%	2.41%	0.41	-0.26	-0.32	-0.23
25	0.70								
26	0.76								
27	0.71								
28	0.67	9.74%	19.14%	*66.87%	3.36%	-0.22	-0.29	0.41	-0.18
29	0.59	8.43%	23.03%	8.81%	*59.02%	-0.27	-0.09	-0.20	0.30
30	0.63	*63.25%	7.81%	20.46%	7.68%	0.40	-0.24	-0.22	-0.21
31	0.73	6.27%	7.53%	*73.29%	12.00%	-0.21	-0.28	0.48	-0.29
32	0.33	*33.22%	24.84%	18.12%	22.96%	0.21	-0.20	-0.04	-0.07
33	0.52								
34	0.68								
35	0.78	6.75%	5.61%	*78.03%	8.18%	-0.31	-0.32	0.53	-0.23
36	0.92	*91.76%	2.36%	2.19%	2.83%	0.52	-0.27	-0.27	-0.31
37	0.88	6.97%	*88.04%	2.17%	1.98%	-0.34	0.52	-0.27	-0.25
38	0.44								
39	0.90	*89.53%	3.90%	2.28%	3.47%	0.39	-0.23	-0.21	-0.21
40	0.79	4.68%	12.28%	*78.96%	3.28%	-0.21	-0.33	0.45	-0.19
41	0.75	15.65%	*74.72%	4.17%	4.45%	-0.25	0.46	-0.27	-0.29
42	0.50								

Table 20 continued									
43	0.76	6.04%	8.52%	8.27%	*75.96%	-0.24	-0.28	-0.26	0.48
44	0.82	*82.24%	10.76%	4.29%	1.65%	0.37	-0.17	-0.33	-0.17
45	0.89	2.76%	*89.18%	3.19%	3.52%	-0.28	0.51	-0.25	-0.30
46	0.71	5.90%	*71.50%	11.02%	10.51%	-0.24	0.52	-0.26	-0.33
47	0.65	*64.73%	5.71%	20.12%	8.29%	0.29	-0.18	-0.19	-0.11
48	0.86	4.12%	*85.67%	2.19%	3.68%	-0.16	0.36	-0.20	-0.22
49	0.70	18.74%	3.76%	*70.10%	6.20%	-0.17	-0.27	0.40	-0.29
50	0.39								

\* indicates the correct answer

Table 21 – HSGQE Mathematics Item Statistics

HSGQE Mathematics Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserials for Each Option			
		A	B	C	D	A	B	C	D
1	0.88	0.98%	8.72%	2.62%	*87.57%	-0.17	-0.18	-0.20	0.26
2	0.76	*76.19%	7.89%	7.68%	7.95%	0.51	-0.34	-0.30	-0.24
3	0.46	*45.55%	1.23%	48.73%	4.07%	0.45	-0.12	-0.40	-0.20
4	0.84	4.51%	7.44%	2.89%	*84.38%	-0.29	-0.34	-0.21	0.49
5	0.83	4.31%	5.90%	*83.24%	6.18%	-0.19	-0.21	0.35	-0.22
6	0.78	4.86%	*77.75%	2.64%	13.74%	-0.19	0.39	-0.19	-0.31
7	0.48								
8	0.68	10.07%	16.92%	*67.74%	3.95%	-0.25	-0.32	0.44	-0.13
9	0.73	1.40%	15.94%	9.82%	*72.54%	-0.16	-0.20	-0.21	0.30
10	0.89	3.70%	5.66%	*88.67%	1.59%	-0.28	-0.24	0.37	-0.14
11	0.84	2.49%	10.39%	2.46%	*84.32%	-0.23	-0.33	-0.23	0.45
12	0.74	3.92%	13.82%	7.74%	*73.95%	-0.26	-0.39	-0.16	0.49
13	0.61	*60.85%	2.89%	32.42%	3.46%	0.44	-0.20	-0.34	-0.25
14	0.56								
15	0.93	*92.60%	3.23%	2.50%	1.20%	0.37	-0.26	-0.21	-0.17
16	0.62	7.86%	11.83%	*62.45%	16.23%	-0.23	-0.27	0.45	-0.24
17	0.72								
18	0.81	2.55%	3.81%	*81.09%	12.22%	-0.20	-0.23	0.32	-0.20
19	0.73	1.57%	18.40%	6.46%	*73.10%	-0.15	-0.40	-0.28	0.52
20	0.74	5.90%	9.75%	9.98%	*73.86%	-0.30	-0.37	-0.26	0.56
21	0.56	12.78%	*56.00%	20.62%	9.55%	-0.33	0.39	-0.10	-0.23
22	0.70	5.97%	*70.31%	22.14%	1.04%	-0.20	0.42	-0.36	-0.14
23	0.60	*60.47%	4.13%	34.12%	0.83%	0.42	-0.17	-0.39	-0.11
24	0.55	7.43%	22.95%	*54.57%	14.49%	-0.32	-0.11	0.46	-0.36
25	0.64	*64.41%	14.32%	12.22%	7.92%	0.56	-0.32	-0.35	-0.22
26	0.25								
27	0.67	25.63%	3.93%	2.94%	*66.93%	-0.36	-0.05	-0.15	0.37
28	0.51	13.98%	*50.76%	9.14%	25.18%	-0.35	0.49	-0.24	-0.18
29	0.62								
30	0.71	4.49%	13.23%	*70.55%	10.93%	-0.18	-0.26	0.41	-0.25
31	0.78	4.60%	5.09%	*77.94%	11.53%	-0.12	-0.25	0.49	-0.41
32	0.53	25.54%	12.78%	7.57%	*52.97%	-0.46	-0.25	-0.09	0.58
33	0.59	3.73%	17.90%	*59.01%	17.67%	-0.11	-0.03	0.25	-0.30
34	0.42	20.82%	24.98%	*41.89%	11.40%	-0.25	-0.13	0.29	-0.06
35	0.48	29.67%	15.13%	*48.31%	5.66%	-0.09	-0.30	0.34	-0.20
36	0.56								
37	0.63	15.22%	*62.69%	12.82%	8.08%	-0.25	0.44	-0.29	-0.16
38	0.78	*77.66%	11.59%	6.94%	2.75%	0.36	-0.30	-0.14	-0.18
39	0.55	28.40%	*54.85%	12.68%	2.62%	-0.40	0.46	-0.16	-0.08
40	0.32	*32.35%	49.28%	8.50%	8.83%	0.29	0.00	-0.28	-0.29
41	0.67	3.58%	18.85%	*67.07%	9.96%	-0.29	-0.15	0.35	-0.24

Table 21 continued									
42	0.94	1.23%	*94.01%	2.46%	1.67%	-0.11	0.31	-0.22	-0.18
43	0.53	9.63%	*53.15%	20.17%	15.82%	-0.15	0.43	-0.32	-0.17
44	0.87	5.45%	2.90%	*86.60%	4.19%	-0.28	-0.25	0.44	-0.23
45	0.64	*64.29%	13.37%	17.84%	3.50%	0.45	-0.44	-0.11	-0.22
46	0.48	29.48%	15.32%	6.20%	*47.67%	-0.17	-0.33	-0.21	0.44
47	0.24					.	.	.	.
48	0.60	7.36%	13.66%	17.62%	*60.06%	-0.25	-0.32	-0.23	0.51
49	0.73	15.23%	*73.14%	7.43%	3.38%	-0.19	0.39	-0.29	-0.23
50	0.46	*45.65%	24.21%	18.92%	10.09%	0.58	-0.18	-0.38	-0.29
51	0.45	24.91%	16.58%	*44.60%	12.58%	-0.13	-0.23	0.37	-0.20
52	0.45	*45.32%	22.10%	22.82%	7.90%	0.31	-0.16	-0.20	-0.12
53	0.67	10.78%	6.24%	*67.46%	14.50%	-0.29	-0.33	0.55	-0.28
54	0.48					.	.	.	.
55	0.41	14.19%	26.01%	18.03%	*40.87%	-0.16	-0.29	-0.18	0.46
56	0.56	9.02%	25.14%	*56.41%	8.59%	-0.29	-0.29	0.50	-0.22
57	0.38	*38.24%	23.44%	23.87%	13.09%	0.45	-0.30	-0.24	-0.04
58	0.55	10.79%	19.52%	*55.11%	13.32%	-0.29	-0.12	0.33	-0.15

\* indicates the correct answer



Table 22 – HSGQE Writing Item Statistics

HSGQE Writing Item Statistics									
ITEM	P-Value	Percent of Total Selecting Each Option				Point Biserials for Each Option			
		A	B	C	D	A	B	C	D
1	0.84	1.28%	4.02%	10.82%	*83.75%	-0.18	-0.25	-0.40	0.47
2	0.78	4.31%	*77.95%	4.07%	13.63%	-0.21	0.35	-0.25	-0.25
3	0.78	5.72%	11.29%	*77.57%	5.32%	-0.24	-0.23	0.35	-0.21
4	0.83	1.91%	3.19%	*83.01%	11.37%	-0.27	-0.26	0.42	-0.30
5	0.57								
6	0.85	5.98%	5.00%	*85.18%	3.60%	-0.26	-0.28	0.43	-0.24
7	0.82	*81.51%	5.14%	6.50%	6.59%	0.37	-0.14	-0.18	-0.35
8	0.63								
9	0.85	9.79%	*84.65%	1.78%	3.56%	-0.28	0.43	-0.27	-0.27
10	0.88	*87.98%	2.84%	2.78%	6.16%	0.39	-0.25	-0.30	-0.20
11	0.80	*80.31%	5.23%	5.92%	8.26%	0.51	-0.31	-0.30	-0.30
12	0.87	5.27%	3.19%	4.45%	*86.72%	-0.22	-0.29	-0.28	0.43
13	0.68								
14	0.67								
15	0.79	15.69%	2.49%	2.24%	*79.09%	-0.23	-0.31	-0.30	0.39
16	0.74	*73.55%	6.05%	7.90%	12.12%	0.40	-0.31	-0.22	-0.20
17	0.84	*84.04%	8.46%	4.01%	2.80%	0.46	-0.29	-0.29	-0.26
18	0.61								
19	0.79	2.49%	7.20%	*78.74%	11.19%	-0.20	-0.33	0.36	-0.17
20	0.79	2.76%	11.06%	4.64%	*78.87%	-0.27	-0.31	-0.31	0.49
21	0.75	2.66%	*74.95%	18.81%	3.19%	-0.22	0.41	-0.30	-0.29
22	0.78	10.07%	7.25%	4.41%	*77.68%	-0.31	-0.31	-0.28	0.52
23	0.64								
24	0.65								
25	0.91	1.84%	5.54%	*91.35%	0.82%	-0.21	-0.34	0.41	-0.13
26	0.85	2.26%	*85.43%	4.67%	7.08%	-0.22	0.38	-0.28	-0.20
27	0.65	*64.80%	19.36%	4.30%	11.01%	0.49	-0.32	-0.34	-0.23
28	0.77	9.31%	9.19%	3.49%	*77.34%	-0.31	-0.31	-0.32	0.53
29	0.89	2.10%	4.78%	*89.04%	3.72%	-0.20	-0.28	0.43	-0.27
30	0.59								
31	0.45								
32	0.82	1.14%	7.58%	8.43%	*81.95%	-0.19	-0.32	-0.24	0.43
33	0.77	12.73%	*76.91%	2.69%	7.19%	-0.25	0.43	-0.23	-0.30
34	0.90	2.45%	4.54%	2.60%	*89.88%	-0.21	-0.27	-0.28	0.43
35	0.59	10.40%	25.91%	4.02%	*59.03%	-0.22	-0.35	-0.31	0.51

\* indicates the correct answer

## Descriptive Statistics and Reliability

Table 23 contains the descriptive statistics and reliability for each grade/content area. The table displays the following statistics for the operational form within a grade/content area:

- the number of scored items;
- the number of score points;
- the case counts (N);
- the raw score means;
- the raw score standard deviations;
- the scale score means;
- the scale score standard deviations

Reliability estimates (measures of internal consistency) are also provided

- Cronbach's  $\alpha$  (Cronbach, 1951)

The reliabilities of the operational forms, as evaluated by Cronbach's  $\alpha$  index of internal consistency, ranged between 0.88 and 0.94. The reliability of the HSGQE Writing test with the weighted score is slightly lower than with the unweighted score (e.g., .0.89 vs. 0.92).

**Table 23 – Descriptive Statistics and Reliability**

Grade	Content Area	No. of Scored Items	Score Points	N	Raw Score		Scale Score		Reliability
					Mean	SD	Mean	SD	Cronbach $\alpha$
3	RD	36	42	9,744	28.76	7.694	357.22	88.85	0.898
	MA	36	44	9,713	30.62	7.992	368.95	86.67	0.887
	WR	36	55	9,732	31.71	9.686	367.91	86.99	0.882
6	RD	36	41	10,491	26.46	7.232	348.40	80.73	0.887
	MA	36	48	10,488	28.97	9.853	354.79	84.06	0.901
	WR	36	58	10,490	36.05	9.884	353.41	85.09	0.897
8	RD	36	41	10,149	28.91	7.700	340.02	88.15	0.891
	MA	36	45	10,088	26.48	8.867	346.97	87.32	0.897
	WR	36	58	10,128	35.87	9.821	341.53	83.43	0.894
HSGQE	RD	50	71	11,121	46.99	13.807	334.77	77.84	0.931
	MA	58	70	11,584	42.44	15.088	349.24	85.88	0.935
	WR*	35	59	10,392	41.25	10.875	338.10	84.50	0.916
	WR+	35	81	10,392	54.77	14.845	338.92	87.40	0.892

\* indicates unweighted

+ indicates weighted

## Percentage of Students in Each Proficient Category

The percentage of students in each proficient category is shown in Tables 24-25.

**Table 24 – Percentage of Students in Each Proficient Category for Benchmark**

Grade Level	Content Area	Not Proficient		Cut Score	Below Proficient		Cut Score	Proficient		Cut Score	Advanced	
		N	%		N	%		N	%		N	%
Grade 3	RD	1262	13	258	1286	13	310	5359	55	433	1844	19
	MA	834	9	254	1905	20	322	3599	37	401	3382	35
	WR	837	9	245	3078	32	352	5145	53	490	679	7
Grade 6	RD	1120	11	248	2052	20	311	3045	29	372	4278	41
	MA	2162	21	291	1584	15	329	3636	35	399	3111	30
	WR	409	4	196	2216	21	300	5448	52	416	2422	23
Grade 8	RD	1156	11	233	917	9	271	1986	20	325	6102	60
	MA	1830	18	272	4340	43	373	3051	30	461	878	9
	WR	383	4	191	3376	33	316	4519	45	416	1861	18

**Table 25 – Percentage of Students in Each Proficient Category for HSGQE**

Administration	Content Area	Not Proficient		Cut Score	Proficient	
		N	%		N	%
Spring 2003	RD	4315	39	322	6808	61
	MA	4409	38	328	7165	62
	WR	2265	22	275	8124	78

## Standard Error of Measurement

An important point to consider when analyzing items/data and interpreting is that each one is a description of a particular performance by the individual or group on the particular test administered. From these descriptions, inferences about the achievement level of the individual or group may be made. The fact that the obtained score for a single test may not represent an individual's true status gives rise to the need for the standard error of measurement (SEM).

Measurement error is associated with every test score. A student's true score is the hypothetical average score that would result if the test could be administered repeatedly without the effects of practice or fatigue. The standard error of measurement can be used to obtain a range within which a student's true score is likely to fall.

An obtained score should be regarded not as an absolute value but as a point within a range that probably includes a student's true score. It is expected that 68% of the time a student's obtained score from a single testing will fall within one SEM of that student's true score, and that 95% of the time the obtained score will fall

within two standard errors of the true score.

Table 26 presents standard errors of measurement and an 80% confidence interval in scale score units for each Benchmark and HSGQE content area. The standard error of measurement is the standard deviation of the distribution of differences between obtained and anticipated scores; the mean difference is zero. The standard error of measurement forms a 68% confidence interval (that is, about two-thirds of the differences will be within one standard error of measurement from the mean). The 80% confidence interval is simply the standard error of measurement multiplied by 1.2816. The difference between a student's obtained and anticipated score is considered significant if it is larger than the 80% confidence interval for the test under consideration.

An example of an applied standard error of measurement would follow this format: student "A" receives a score of 415 in the content area of Reading at Grade 3. One could state with a 68% degree of reliability that student A's score in the Reading content area would fall within the range of 377 and 453 no matter how many times student A took the test (an 80% reliability would fall within the range of 366 and 464).

In the formula below, used to compute the standard error of measurement,  $R$  is the test reliability coefficient and  $\sigma$  is the standard deviation in scale score units on which  $R$  was computed:

$$SE_{\text{measurement}} = \sigma_{ss} \sqrt{1 - R^2}.$$

Table 27 present the SEM in raw score units, and is compared to TerraNova. As shown in Table 27, TerraNova has smaller SEM values than the Alaska tests.

**Table 26 – Scale Score SEM's and 80% Confidence Intervals**

<b>Content Area</b>	<b>SE<sub>measurement</sub></b>	<b>80% CI</b>
<b>Grade 3</b>		
Reading	39.09	50.10
Mathematics	40.02	51.29
Writing	40.99	52.54
<b>Grade 6</b>		
Reading	37.28	47.78
Mathematics	36.47	46.74
Writing	37.61	48.20
<b>Grade 8</b>		
Reading	40.02	51.29
Mathematics	38.60	49.47
Writing	37.38	47.91
<b>HSGQE</b>		
Reading	28.41	36.41
Mathematics	30.46	39.03
Writing*	33.90	43.45
Writing+	39.51	50.63

\* indicates unweighted

+ indicates weighted

**Table 27 – Raw Score Based SEM Comparison to TerraNova for 2002**

	<b>Spring 2003</b>		<i><b>TerraNova (Multiple Assessments Form A)</b></i>		
Grade/Content	SEM	80%CI	SEM	80% CI	Level
Gr3 RD	3.39	4.34	3.14	4.02	13
Gr3 MA	3.69	4.73	3.28	4.20	13
Gr3 Wr	4.56	5.85	2.58	3.31	13
Gr6 RD	3.34	4.28	3.63	4.65	16
Gr6 MA	4.27	5.48	3.88	4.97	16
Gr6 Wr	4.37	5.60	3.32	4.25	16
Gr8 RD	3.50	4.48	3.37	4.32	18
Gr8 MA	3.92	5.02	3.45	4.42	18
Gr8 Wr	4.40	5.64	3.33	4.27	18
HS RD	5.04	6.46	3.20	4.10	20
HS MA	5.35	6.86	4.00	5.13	20
HS WR*	4.36	5.59	3.68	4.72	20
HS WR+	6.71	8.60	N/A	N/A	20

\* indicates items were unweighted

+ indicates items were doubled weighted

### Inter-Rater Reliability

The 2003 operational examination required double readings for all of the HSGQE student responses to each constructed response (CR) item, whereas the Benchmark student responses to the CR items had double reads for 20% of them. The Benchmark examinations used single reads to obtain scores on each CR item, both two-point and three-point short response (SR) items, and four-point or higher extended response (ER) items. For the two-point to higher extended response, approximately 20% of the sampled students obtained additional readings. Every fifth paper in the data collected was read a second time.

Agreement rates for the CR items are presented along with the correlation of the first and second judges' ratings for both the HSGQE and Benchmark double reads. These agreement rates provide useful information about the inter-rater reliability (i.e., single reads with a 20% sample obtaining multiple readings for Benchmark tests and the double reads for the HSGQE).

Readers were trained to implement the scoring rubrics, and anchor papers, check sets, and read-behinds were employed. Item scores for CR items were obtained on the basis of single readings of student responses and used in both the scoring of students and the scaling of the HSGQE and Benchmark forms. If the two readings for a two-point or three-point SR item differed, a third independent reading was obtained. A third reading of a student response to a four-point or higher ER item was required if the two readings differed by more than one point.

Tables 28 - 39 give complete Rater Analyses of each grade level and content area. The items listed are the items that provided a point range for scoring purposes. Total exact agreement refers to the percentage of times the raters agreed completely, the "1pt. Difference" is the percentage of time when raters disagreed by a margin of 1 point in either direction, and the "2+ pt. Difference" refers to the percentage of time when raters disagreed by 2 or more points in either direction. The percent-missing columns are those students that did not supply an answer to the item being scored.

**Table 28 – Benchmark 1 Mathematics Rater Analyses**

Benchmark 1 Mathematics Grade 3 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
6	94.96	3.26	0.25	98.47	1.53	100.00
7	88.96	7.53	1.63	98.12	1.88	100.00
17	89.67	7.68	0.10	97.46	2.54	100.00
18	92.17	4.68	0.56	97.41	2.59	100.00
29	92.88	3.87	0.00	96.74	3.26	100.00
30	93.49	3.05	0.00	96.54	3.46	100.00



**Table 29 – Benchmark 1 Reading Rater Analyses**

Benchmark 1 Reading Grade 3 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
15	86.16	9.41	1.78	97.36	2.64	100.00
16	92.62	1.63	0.00	94.25	5.75	100.00
17	90.79	4.73	0.00	95.52	4.48	100.00
22	92.57	1.63	0.05	94.25	5.75	100.00
33	80.62	13.53	0.15	94.30	5.70	100.00
36	86.42	7.58	0.10	94.10	5.90	100.00

**Table 30 – Benchmark 1 Writing Rater Analyses**

Benchmark 1 Writing Grade 3 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
9	63.16	31.30	1.37	95.83	4.17	100.00
12	83.97	8.75	0.41	93.13	6.87	100.00
17	64.38	30.84	0.61	95.83	4.17	100.00
25	78.42	9.87	0.10	88.40	11.60	100.00
26	94.76	1.93	0.00	96.69	3.31	100.00
32	69.62	25.34	1.32	96.28	3.72	100.00

**Table 31 – Benchmark 2 Mathematics Rater Analyses**

Benchmark 2 Mathematics Grade 6 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
10	94.03	2.12	0.61	96.76	3.24	100.00
11	92.52	4.56	0.19	97.27	2.73	100.00
19	83.31	13.63	0.61	97.56	2.44	100.00
20	82.79	13.02	0.61	96.43	3.57	100.00
28	93.47	4.94	0.05	98.45	1.55	100.00
29	84.11	6.86	0.05	91.02	8.98	100.00
32	70.95	19.32	3.95	94.22	5.78	100.00
33	94.12	2.96	0.00	97.09	2.91	100.00

**Table 32 – Benchmark 2 Reading Rater Analyses**

Benchmark 2 Reading Grade 6 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
9	92.10	4.89	0.09	97.09	2.91	100.00
14	94.88	3.10	0.00	97.98	2.02	100.00
25	76.82	18.81	0.94	96.57	3.43	100.00
34	87.82	5.36	0.05	93.23	6.77	100.00
35	86.32	9.87	0.00	96.19	3.81	100.00
36	81.15	14.48	0.00	95.63	4.37	100.00

**Table 33 – Benchmark 2 Writing Rater Analyses**

Benchmark 2 Writing Grade 6 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
5	49.93	41.28	6.72	97.93	2.07	100.00
9	65.73	31.03	0.42	97.18	2.82	100.00
13	60.60	32.21	0.85	93.65	6.35	100.00
19	81.52	13.40	0.75	95.67	4.33	100.00
29	58.44	34.60	1.13	94.17	5.83	100.00
35	89.84	7.43	0.00	97.27	2.73	100.00
36	64.60	30.79	0.75	96.14	3.86	100.00

**Table 34 – Benchmark 3 Mathematics Rater Analyses**

Benchmark 3 Mathematics Grade 8 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
11	83.87	12.37	0.10	96.34	3.66	100.00
21	83.68	10.88	0.10	94.66	5.34	100.00
22	85.51	10.93	0.67	97.11	2.89	100.00
23	85.85	7.08	0.14	93.07	6.93	100.00
30	86.33	4.72	0.29	91.33	8.67	100.00
33	84.69	2.21	0.14	87.05	12.95	100.00
36	90.52	3.32	0.14	93.98	6.02	100.00

**Table 35 – Benchmark 3 Reading Rater Analyses**

Benchmark 3 Reading Grade 8 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
10	77.18	17.81	1.01	96.00	4.00	100.00
12	72.56	20.94	0.87	94.37	5.63	100.00
13	83.34	6.64	0.00	89.99	10.01	100.00
26	94.80	0.91	0.00	95.72	4.29	100.00
29	81.32	12.90	0.87	95.09	4.91	100.00
33	75.93	16.75	0.72	93.40	6.60	100.00

**Table 36 – Benchmark 3 Writing Rater Analyses**

Benchmark 3 Writing Grade 8 Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
6	60.28	33.51	1.88	95.67	4.33	100.00
10	89.55	5.63	0.29	95.47	4.53	100.00
12	62.69	32.31	0.39	95.38	4.62	100.00
18	63.41	31.30	0.53	95.23	4.77	100.00
29	63.60	29.32	0.58	93.50	6.50	100.00
36	58.35	32.11	0.96	91.43	8.57	100.00

**Table 37 – HSGQE Mathematics Rater Analyses**

HSGQE Mathematics Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
7	77.08	3.53	0.27	80.88	19.12	100.00
14	78.72	4.52	0.02	83.26	16.74	100.00
17	79.78	2.34	0.37	82.49	17.51	100.00
26	76.39	1.29	0.08	77.76	22.24	100.00
29	76.77	5.39	0.50	82.66	17.34	100.00
36	66.26	9.55	0.38	76.19	23.81	100.00
47	67.92	2.61	0.07	70.59	29.41	100.00
54	71.89	3.39	0.12	75.41	24.59	100.00

**Table 38 – HSGQE Reading Rater Analyses**

HSGQE Reading Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
4	78.12	7.60	0.22	85.94	14.06	100.00
5	72.06	11.70	0.58	84.34	15.66	100.00
10	74.18	11.00	0.96	86.13	13.87	100.00
16	77.63	6.46	0.16	84.25	15.75	100.00
18	73.43	4.60	0.11	78.14	21.86	100.00
23	69.72	13.67	0.49	83.89	16.11	100.00
25	63.70	17.58	0.94	82.22	17.78	100.00
26	76.65	6.19	0.19	83.04	16.96	100.00
27	66.95	14.20	0.78	81.93	18.07	100.00
33	63.59	14.97	0.92	79.48	20.52	100.00
34	63.17	15.55	4.01	82.74	17.26	100.00
38	68.20	13.03	0.62	81.85	18.15	100.00
42	80.12	4.49	0.08	84.69	15.31	100.00
50	77.12	6.32	0.08	83.52	16.48	100.00

**Table 39 – HSGQE Writing Rater Analyses**

HSGQE Writing Form G Spring 2003 Rater Analysis						
Item Number	Total Exact Agreement	1 pt. Difference	2+ pt. Difference	Total Percentage Rated	Percentage Missing	Total Rated + Missing
5	48.98	29.90	1.56	80.44	19.56	100.00
8	79.56	5.67	0.00	85.23	14.77	100.00
13	77.37	6.83	0.22	84.41	15.59	100.00
14	53.71	29.48	0.74	83.93	16.07	100.00
18	51.87	28.27	0.77	80.92	19.08	100.00
23	77.46	6.29	0.15	83.89	16.11	100.00
24	53.73	27.72	0.55	81.99	18.01	100.00
30	49.09	29.34	0.82	79.25	20.75	100.00
31	63.90	16.36	2.05	82.31	17.69	100.00

**Table 40 – Exact Agreement Rates by Grade/Content Area**

Grade	Reading		Mathematics		Writing	
	Smallest Exact Agreement	Largest Exact Agreement	Smallest Exact Agreement	Largest Exact Agreement	Smallest Exact Agreement	Largest Exact Agreement
3	80.62	92.62	88.96	94.96	63.16	94.76
6	76.82	94.88	70.95	94.12	49.93	89.84
8	72.56	94.80	83.68	90.52	58.35	89.55
HSGQE	63.17	80.12	66.26	79.78	48.98	79.56

**Table 41 – Average Agreement Rates by Grade/Content Area**

Content Area	Defined Measurement	Grade			
		3	6	8	HSGQE
Reading	Exact	88.20	86.52	80.86	71.76
	Adjacent	6.42	9.42	12.66	10.53
	Total	94.62	95.94	93.52	82.29
Math	Exact	92.02	86.91	85.78	74.35
	Adjacent	5.01	8.43	7.36	4.08
	Total	97.03	95.34	93.14	78.43
Writing	Exact	75.72	67.24	66.31	61.74
	Adjacent	18.01	27.25	27.36	19.98
	Total	93.73	94.49	93.67	81.72

CR item agreement rates, both exact (i.e., identical readings for the first two readings) and adjacent (the two readings were within one point of each other), were calculated. Total agreement rates were also calculated by adding the exact and approximate rates together. Table 40 shows the smallest and largest exact agreement rates by grade/content area. As shown, the smallest exact agreement rate was 49.93% for Grade 6 Writing and the largest exact agreement rate was 94.96% for Grade 3 Mathematics.

Table 41 provides each grade/content area averages within exact and adjacent figures given. The total agreement rates provided are the summed exact and approximate figures. The total agreement rates range from 78.43% to 95.94%, showing that inter-rater reliability was high for the most part. Total agreement rates tended to be highest for SR items. The very high total agreement rates observed for these items is expected, given that each of these items has a maximum value of three points.

### Calibration and Equating

CTB has recalibrated the Spring 2003 operational form of the HSGQE for each content area. This recalibration was necessary because of possible context effects of item parameters that were used from the field test. Such context effects could make items in the operational forms easier or harder than they were in the field test due to changes in the location or context in which the items are presented from the field test administration to the operational test administration.

To address these effects, CTB implemented a recalibration of an operational form at the time of its administration, and a subsequent linking of the operational form back to the scale developed at field test using the procedure defined by Stocking and Lord (1983). The Stocking and Lord procedure can be implemented easily using PARDUX.

### Calibration

The multiple-choice and open-ended items were calibrated together for the operational test. This was done in part because a single scale that reflects the trait assessed by the two item types is theoretically attractive and technically feasible. The single scale also communicates the instructionally sound idea that the skills to be assessed relate to the same underlying qualities and characteristics, and that they can be taught and measured using a variety of assessment modes. In considering the simultaneous calibration process, it is also important to recall the position of D. Thissen, H. Wainer, and X-B. Wang (1992), that items of diverse types can be scaled together provided the different types of items assess the same primary characteristics.

### 3PL/2PPC Models

The item response theory (IRT) calibration models used for scaling MC and CR items together was the three-parameter logistic (3PL) model to scale the MC items and the two-parameter, partial credit (2PPC) model (Yen, 1981) to scale the CR items. A brief explanation of the models is provided below.

It is important to note that although the one-parameter model is sometimes used for scaling, it was not considered for this test. CTB typically does not recommend use of the one-parameter model because it is unduly restrictive, permitting items to vary only in terms of difficulty; it is our experience that MC items also vary in terms of discrimination, and that examinees' correct answers to these items may reflect guessing. As Divgi (1986) noted, MC items are unlikely to meet the assumptions of the one-parameter model.

The 3PL model (Lord, 1980) was used in the analysis of MC items. In this model, the probability that a student with an IRT ability estimate  $\theta$  responds correctly to item  $i$  is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

where  $a_i$  is the item discrimination,  $b_i$  is the item difficulty, and  $c_i$  is the probability of a correct response by a very low-scoring student.

For analysis of the CR items, a 2PPC model was used. This model is a special case of Bock's nominal model (1972). Bock's model states that the probability of an examinee with ability  $\theta$  having a score at the  $k$ -th level of the  $j$ -th item is

$$P_{jk}(\theta) = P(x_j = k - 1 | \theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \quad k = 1 \dots m_j,$$

where

$$Z_{jk} = A_{jk} \theta + C_{jk}.$$

For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j(k - 1),$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \quad \text{where } \gamma_{j0} = 0,$$

where  $A_j$  and  $\gamma_{ji}$  are the estimated parameters.

The first constraint implies that higher item scores reflect higher ability levels and that items can vary in their discriminations. For the 2PPC model, each item has  $m_j - 1$  independent  $\gamma_{ji}$  parameters and one  $a_j$  parameter. A total of  $m_j$  independent item parameters are estimated.

Table 42 presents the summary of calibration results for the individual content areas of the 2003 Benchmark and the HSGQE Assessments.

**Table 42 – Summary of Calibration Results**

Grade	Content Area	No. of Items	Sample Size (observed)	Raw Score		Max A	Default C	B Value Range	No. of Est. Cycles	Non – Conv Items
				Mean	SD					
3	RD	36	9,744	28.76	7.694	1	5	-4.384 to 3.166	44	0
	MA	36	9,713	30.62	7.992	0	4	-3.066 to 1.571	37	0
	WR	36	9,732	31.71	9.686	0	1	-4.301 to 2.072	37	0
6	RD	36	10,491	26.46	7.232	0	4	-4.791 to 1.690	37	0
	MA	36	10,488	28.97	9.853	0	5	-3.116 to 0.561	23	0
	WR	36	10,490	36.05	9.884	0	8	-2.719 to 2.234	18	0
8	RD	36	10,149	28.91	7.700	0	9	-3.646 to 0.832	31	0
	MA	36	10,088	26.48	8.867	0	3	-3.543 to 2.044	25	0
	WR	36	10,128	35.87	9.821	0	6	-4.170 to 2.381	26	0
HSGQE	RD	50	11,121	46.99	13.807	1	4	-4.261 to 3.010	25	0
	MA	58	11,584	42.44	15.088	0	7	-3.501 to 1.639	10	0
	WR	35	10,392	41.25	10.875	0	4	-3.143 to -0.143	12	0

Note: Max A = maximum a – parameters (discrimination); Default C = c – parameters (guessing) value is 0.25;  
B Value Range = b – parameters (difficulty); No. of Est. Cycles = number of estimation cycles

### Item Fit and Nonconvergence

Occasionally, a test may contain misfit items. These misfit items were flagged using a set criterion. A procedure described by Yen (1981) was used to measure fit to the 3PL (multiple choice), 2PPC (constructed-response) model (see Appendix A). Table 43 shows the number of items by grade and content area that were considered to have misfit. On the average across all grades/content areas, about 5% of the questions fit the model poorly. High school Writing had the highest percentage (17.14%) of misfit items, while Grade 6 Reading had the lowest percentage (0%). With the exception of these extreme cases, there was no noticeable difference between grades or content areas in terms of the percentage of misfit items.

Currently, numbers of misfit items are not available in *TerraNova*. There were some misfit items in the tryout. However, when the final standardized version of *TerraNova* was developed, misfit items were eliminated.

There were no items across all grades and content areas that would not converge using the set criterion.



**Table 43 – Number of Misfit Items**

Grade	Content Area	Total # of Items	Number of Misfit Items	Percentage of Misfit Items
3	RD	36	2	5.56
	MA	36	2	5.56
	WR	36	1	2.78
6	RD	36	0	0.00
	MA	36	1	2.78
	WR	36	5	13.89
8	RD	36	2	5.56
	MA	36	1	2.78
	WR	36	3	8.33
HSGQE	RD	50	3	6.00
	MA	58	1	1.72
	WR	35	6	17.14

There were no items with collapsed score levels for the operational forms. An item results in a collapsed score level when the maximum number of points for that item is not achieved. If items were found to have collapsed scores levels, these items would be reviewed by the development and scoring teams to determine if the maximum score point was viable. In most cases, it would be determined that the maximum score point met reasonable expectations.

### Dimensionality

One important dimensionality issue was whether mixed item types should be scaled together or apart. If the mixed item types did not measure a single dimension, the CR items and MC items would have been scaled separately. The purpose of these analyses was to determine whether the HSGQE and Benchmark items within a grade/content area assessed the same ability, and hence, whether the item types could continue to be scaled together.

The Q3 statistic (Yen, 1984) was used for this evaluation. It was obtained by correlating differences between students observed and expected responses for pairs of items after taking into account overall test performance. If a substantial number of items in the test demonstrate local dependence, these items may be calibrated separately. Pairs of items with Q3 values greater than 0.20 were classified as locally dependent (Yen, 1984). The maximum value for this index is 1.00. The content of the flagged items was examined in order to identify possible sources of the local dependence.

The number of item pairs flagged under the criterion varied across forms and content areas. In most cases, the dependence of the questions was probably due to the two questions involving a similar task (i.e., “Write a paragraph about...”) or both questions relating to the same item.

There were only a few cases where the dependency was undetermined. Table 44 below shows the total number of pairs flagged by grade/content area. It is apparent that there was a higher frequency of dependent items in the Writing content area compared to Reading and Mathematics. This was most likely due to similarity in the Writing items.

**Table 44 – Item Pair Dependence by Grade/Content Area**

Content Area	Grade Level			
	3	6	8	HSGQE
Reading	1	0	1	1
Mathematics	1	0	1	1
Writing	1	4	5	10
Numbers of item pairs are listed				

In conclusion, it appears that dependence could be attributed to items manifesting passage/item dependence and requiring similar tasks. Upon inspection of the content of the item pairs, it can be seen that no systematic patterns of dependence could be attributed to item format. Overall, the items exhibiting dependency were not of sufficient magnitude to warrant concern. Analyses will be conducted to monitor the presence of local dependence in future forms of the test.

#### Setting the Scale Units and Values for the LOSS and HOSS

The LOSS (Lowest Obtainable Scale Score) and HOSS (Highest Obtainable Scale Score) were set for the purpose of an operational scale. These values were established based on an examination of the scale score distributions and standard error (SE) curves. All grade/content areas had the same LOSS and HOSS values. In each case, the LOSS was 100 and the HOSS was 600.

#### Equating: The 2003 Operational Test Scale

The linking between the 2002 scale and the 2003 administration was accomplished through the use of anchor items (anchors items are items that remain the same textually and in the same location from operational form to operational form). We also used Stocking and Lord (1983) procedures to transform the operational score scale. The Stocking and Lord procedure finds a linear transformation that best aligns the characteristic curves defined by the common anchor items from the 2002 and 2003 operational tests.

Table 45 summarizes the results of the Equating process.

**Table 45 – Summary of Equated Item Parameters for HSGQE**

Content Area	No. of Anchors	P-value Comparison After Equating			
		Diff.	RMSD	SD Ratio	r
Reading	17	0.010	0.049	1.015	0.954
Mathematics	20	0.001	0.034	0.963	0.970
Writing	11	0.006	0.049	0.907	0.861

Note: No equating is done for Benchmark since the items are the same across administrations.

## Bias Studies

CTB minimizes the presence of potentially biased items by instituting a strict set of procedures which include in-house reviews, external bias screening committees, and statistical differential item functioning (DIF) analysis. Using the judgment of a bias review committee is an essential step in ensuring that all visible sources of bias or potential bias are eliminated from stimulus materials, items, and artwork prior to field testing items.

## Bias Reviews

Four procedures were used to reduce bias in the Alaska HSGQE and the Benchmark. The first is based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias can occur only if the test is measuring different things for different groups. If the test entails irrelevant skills or knowledge (however common), the possibility of bias is increased. Thus, careful attention was paid to content validity, doing much to eliminate the possibility of bias.

The second step required the use of the CTB/McGraw-Hill guidelines designed to reduce or eliminate bias. Item writers were given specifications that include directions to adhere to the following materials: *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (MacMillan/McGraw-Hill, 1993) and *Guidelines for Bias-Free Publishing* (McGraw-Hill, 1983). Editors' review test materials with these considerations in mind. These internal editorial reviews of field test materials were done by at least four separate people: the content editor, who directly supervises the item writers; the project director; a style editor; and a proofreader.

In the third procedure, Alaskan educational community professionals who represent various ethnic groups reviewed all field test materials. These reviewers are asked to consider and comment on the appropriateness of language, subject matter, and representation of people.

The bias reviews conducted by panelists are supplemented with statistical procedures carried out during the

data analysis from the field test. The current evidence suggests that expertise in this area is no substitute for data. Reviewers are often incorrect about which items work to the disadvantage of a group, apparently because some of their ideas about student reaction to items are faulty (Sandoval and Mille, 1979; Jensen, 1980; Scheuneman, 1984). For subgroups of the population where known characteristics such as gender and ethnicity may be to the advantage or disadvantage of members of the subgroups, differential item functioning (DIF) studies were performed. The DIF studies were also performed based on regional and community type classifications. The primary source of data describing the regional and community demographic characteristics for DIF analyses was supplied by the ADEED. Each of the districts was put into one of five geographic regions and placed into one of three community types as indicated in Tables 46 and 47, respectively.

**Table 46 – Alaskan Regional Classifications**

Number	Regional Names
1	Interior Region
2, 3, 4 <sup>1</sup>	Northwest Arctic Borough, North Slope Borough Region, Norton Sound Region (a.k.a. North Arctic)
5	Southeast Region
6	Southern Region
7	Yukon Kuskokwim Region

<sup>1</sup> Regions 2 through 4 were combined into one region (North Arctic).

**Table 47 – Alaskan Classification and Criteria for Community Type**

Number	Community Type	Criteria
1	Urban	More than 2500 students located in a single urban center.
2	Rural	Small single-site districts or large districts consisting of many small scattered schools, the majority of which are connected to a road system, the Alaska Marine Highway System, or receive daily service from Alaska Airlines jets (737 or larger).
3	Remote	Districts that are not connected to a road or the Alaska Marine Highway System (state ferry system) or do not receive daily service from Alaska Airlines jets (737 or larger).

## Differential Item Functioning

Differential item functioning (DIF) is a difference in item performance between two groups after controlling examinees' overall achievement level. DIF was evaluated for the operational test items using two procedures: Linn and Harnisch (L-H) and Standardized Mean Difference (SMD). The entire population of students participating in the Spring 2003 testing was used for all analyses. The following subgroups were analyzed:

1. African Americans
2. Caucasians
3. Hispanics
4. Alaskan Natives
5. Asian-Pacific Islanders
6. Females (regardless of racial/ethnic group)
7. Males (regardless of racial/ethnic group)
8. LEP
9. Special Education

The Linn and Harnisch (1981) procedure was implemented because it utilizes the predictions of the three-parameter model (and has been extended to the two-parameter partial credit model), but does not require as many cases as other IRT-based procedures (Lord, 1980 and Linn, Levine, Hasting, & Wardrop, 1981). The L-H statistics, as implemented by the PARDUX program, allowed evaluations within gender groups and within ethnic groups, but not between the combinations.

The SMD was implemented because it takes into account the natural ordering of the response levels of the items and has the desirable property of being based on those ability levels where members of the focal group are present. DIF was evaluated between gender, across Caucasian, African-American, and Hispanic ethnic groups, and across combinations of these gender and ethnic groups for the SMD statistics. Two additional analyses were conducted to investigate the DIF associated with community type and regional assignment. For the last investigation, the following subgroups were analyzed:

1. Community Type
  - a) Rural vs. Urban
  - b) Remote vs. Urban
2. Regional
  - a) Interior vs. Southern
  - b) North Arctic vs. Southern
  - c) South Eastern vs. Southern
  - d) Yukon vs. Southern

*Linn and Harnisch*

To assess DIF for the Benchmark and HSGQE items, item responses for students identified as African American, Hispanic, Alaskan Native, or Asian-Pacific Islander were examined, as were responses for the remainder (Reference Group). Gender analyses were also conducted. The Linn-Harnisch output from PARDUX provided information on the systematic differences between the obtained and expected frequencies. An example of this procedure for ethnic DIF analyses follows for the multiple-choice items.

The parameters for each multiple choice item ( $a_i$ ,  $b_i$ , and  $c_i$ ) and the ability or scale score estimate ( $\theta$ ) for each examinee are estimated for the three-parameter logistic model

$$P_{ij} = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i (\theta_j - b_i)]},$$

where  $P_{ij}$  is the probability that examinee  $j$  will pass item  $i$ . Note that the item parameters are based on item responses from the total sample of examinees, which includes all categories (African-American, Hispanic, Alaskan Native, Asian-Pacific Islander, and Reference Group) in the operational test samples. The sample is then divided into ethnic groups, and the members of each group are sorted into ten equal score categories (deciles) based upon their location on the scale score ( $\theta$ ) scale. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group.

The proportion of people in decile  $g$  who are expected to answer item  $i$  correctly is

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij},$$

where  $n_g$  is the number of examinees in decile  $g$ . The proportion of people expected to answer item  $i$  correctly (over all deciles) for a group (e.g., African American) is

$$P_{i\cdot} = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile ( $O_{ig}$ ) is the number of examinees in decile  $g$  who answered item  $i$  correctly divided by the number of people in the decile ( $n_g$ ). That is,

$$O_{ig} = \frac{\sum_{j \in g} u_{ij}}{n_g},$$

where  $u_{ij}$  is the dichotomous score for item  $i$  for examinee  $j$ .

The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete ethnic group is given by

$$O_{i\cdot} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g} .$$

After the values are calculated for these variables, the difference between the observed proportion correct (for an ethnic group) and expected proportion correct can be computed. The decile group difference ( $D_{ig}$ ) for the observed and expected proportion correctly answering item  $i$  in decile  $g$  is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference ( $D_{i\cdot}$ ) between observed and expected proportion correct for item  $i$  in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot} .$$

These indices are indicators of the degree to which members of an ethnic group perform better or worse than expected on each item, based on the parameter estimates from all subsamples. Differences for decile groups provide an index for each of the ten regions on the scale score ( $\theta$ ) scale. The decile group difference ( $D_{ig}$ ) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of  $\theta$  and a large negative difference in another range of  $\theta$ , yet have a small overall difference.

Items are flagged as demonstrating DIF for or against the specified subgroup according to the following rule. An item demonstrates DIF against a subgroup if the value of the obtained mean minus the expected mean is  $\leq -0.10$ , and the corresponding  $Z$  value is  $\geq 2.58$ . DIF in favor of a subgroup is defined in the same way but with a positive difference.

Table 48 shows the number of significant items with DIF, and it is apparent that very few items demonstrated DIF for or against the subgroups (either gender or ethnicity). Each of the DIF analyses for the subgroups were conducted independently from other subgroups. Items flagged for DIF could be the same item flagged in each subgroup.

Looking across all grades/content areas, more items displayed DIF between ethnic groups than genders, and Asian-Pacific Islanders seemed the most active in the degree and number of items showing either positive or negative DIF. In addition, the number of items with DIF stayed fairly consistent across content areas in Grades 3, 6, 8, and HSGQE.

**Table 48 – Number of Significant DIF – Linn-Harnisch Statistics**

Grade/ Content Area	Male		Female		African-American		Alaskan Native		Asian-Pacific Islander		Hispanic		Caucasian	
	Number of Items with Either Negative or Positive Bias													
	-	+	-	+	-	+	-	+	-	+	-	+	-	+
3/RD	0	0	0	0	0	1	0	0	1	0	0	0	0	0
3/MA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3/WR	1	0	0	1	0	0	0	0	0	1	0	0	0	0
6/RD	0	0	0	0	0	1	0	0	0	0	0	0	0	0
6/MA	1	0	0	1	0	0	1	0	1	1	0	0	0	0
6/WR	1	0	0	1	0	0	0	0	2	4	0	0	0	0
8/RD	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8/MA	0	0	0	0	0	0	0	0	0	1	0	0	0	0
8/WR	1	0	0	0	0	0	0	2	0	1	0	1	0	0
HS/RD	0	0	0	0	2	1	0	1	0	1	0	0	0	0
HS/MA	1	0	0	1	0	1	1	0	1	3	0	0	0	0
HS/WR	2	0	0	2	2	0	0	1	2	4	0	0	0	0

Note: Very small sample size for American Indian, therefore no data presented.



### *Standardized Mean Difference*

The standardized mean difference (SMD) statistic (Zwick et. al., 1993) provides an acceptable alternative to the Mantel-Haenszel statistics when used on tests with polytomously as well as dichotomously scored items. The SMD statistics can provide DIF information for both dichotomous and polytomous items, whereas a single Mantel-Haenszel odds ratio estimator is not available for polytomous items. The SMD takes into account the natural ordering of the response levels of the items and has the desirable property of being based on those ability levels where members of the focal group are present. The SMD output results in a single statistic for each item. An example of this procedure for ethnic DIF analyses follows

$$SMD = \sum p_{Fk} m_{Fk} - \sum p_{Rk} m_{Rk},$$

where  $p_{Fk} = n_{F+k}/n_{F++}$  is the proportion of focal group members who are at the  $k$ th level of the matching variable,

$m_{Fk} = (1/n_{F+k})(\sum y_i n_{Rik})$  is the mean item score for the focal group at the  $k$ th level, and

$m_{Rk} = (1/n_{R+k})(\sum y_i n_{Rik})$  is the analogous value for the reference group. Table 49 lists the criteria used for each individual item.

**Table 49 – DIF Rating Criteria**

Criteria	DIF Rating	Meaning for a Focal Group
If: value $\leq -0.20$	-2	Large unfavorable bias
If: $-0.19 \leq \text{value} \leq -0.10$	-1	Moderate unfavorable bias
If: $-0.09 \leq \text{value} \leq +0.09$	0	Not significant
If: $+0.10 \leq \text{value} \leq +0.19$	1	Moderate favorable bias
If: value $\geq +0.20$	2	Large favorable bias

DIF ratings for each of the variables are described as follows

- ◆ -2 and -1, value means large and moderate amounts of DIF against a focal group
- ◆ 0, implying no DIF against the focal group
- ◆ +1 and +2, value means moderate and large amounts of DIF in favor of a focal group.

A moderate amount of practically significant DIF, for or against the focal group, is represented by an SMD with an absolute value between 0.10 and 0.19, inclusive. A large amount of practically significant DIF is represented by an SMD with an absolute value of 0.20 or greater.

Tables 50 – 53 contain a summary of the number of items that have been flagged for each of the comparison groups for a moderate or large amount of practically significant DIF. Both types of practically significant DIF were associated with significant Mantel statistics ( $\alpha = 0.05$ ) in previous research (Michigan HSPT: Field Test

Racial and Gender Bias Analyses). The tables show SMD summary results for the HSGQE and Benchmark grades are broken down by item type. A DIF rating of “0” means that there was no bias. Across all grades/content areas and various subgroups, the majority of items had a DIF rating of “0.” The number of score points for an item did not affect the DIF rating.

DIF rating summary results for special populations require a larger sample of students in each subgroup analyzed to produce valid information. For this reason, these results would be considered psychometrically questionable. The special populations include students with accommodations, special education students, and LEP students and are presented under the Special Programs section of the tables.

**Table 50 – Summary of Measured DIF – Item Summary Table – HSGQE**

Number of items of the type showing the amount of DIF for the column. For DIF Scale, see DIF Rating Criteria chart.																
Categories of Comparisons		Multiple Choice Items and Constructed Response Items Separately Compared														
		Number of Items with each value on the DIF Scale														
		Math Items = 58					Reading Items = 50					Writing Items = 35				
DIF Scale =		2	1	0	-1	-2	2	1	0	-1	-2	2	1	0	-1	-2
<b>Regions</b> Interior.....1 North Arctic.....2 Southeast.....5 Southern.....6 Yukon/Kus.....7	6 to 1	0	0	57	1	0	0	0	50	0	0	0	0	34	1	0
	6 to 2	0	0	55	2	1	1	1	47	1	0	0	5	28	2	0
	6 to 5	0	1	56	0	1	0	0	50	0	0	0	0	35	0	0
	6 to 7	0	1	53	3	1	1	2	43	4	0	0	6	22	6	1
<b>Communities</b> Urban.....1 Rural.....2 Remote.....3	1 to 2	0	0	57	1	0	0	0	50	0	0	0	0	35	0	0
	1 to 3	0	0	56	1	1	1	0	47	2	0	0	4	31	0	0
<b>Culture / Ethnicity</b> Alaskan Native...0 Native Amer.....1 Asian.....2 Afr. Amer.....3 Hispanic Amer...4 Caucasian.....5	5 to 0	0	0	57	1	0	1	1	46	2	0	0	4	29	2	0
	5 to 1	0	1	57	0	0	0	1	49	0	0	0	2	30	2	1
	5 to 2	1	3	53	1	0	1	3	45	1	0	2	2	27	3	1
	5 to 3	1	0	56	1	0	0	3	45	2	0	0	2	31	2	0
	5 to 4	0	0	58	0	0	0	1	49	0	0	0	1	33	1	0
<b>Gender</b> Male and Female	M to F	1	0	57	0	0	0	5	43	2	0	2	1	31	1	0
<b>Special Programs</b> Regular.....1 Special Edu.....2 LEP.....3	1 to 2	0	0	58	0	0	0	0	50	0	0	0	1	33	1	0
	1 to 3	0	0	57	1	0	1	1	47	1	0	2	4	23	5	1

**Table 51 – Summary of Measured DIF – Item Summary Table – Benchmark 1**

Number of items of the type showing the amount of DIF for the column. For DIF Scale, see DIF Rating Criteria chart.																
Categories of Comparisons		Multiple Choice Items and Constructed Response Items Separately Compared														
		Number of Items with each value on the DIF Scale														
		Math Items = 36					Reading Items = 36					Writing Items = 36				
DIF Scale =		2	1	0	-1	-2	2	1	0	-1	-2	2	1	0	-1	-2
<b>Regions</b>	6 to 1	0	0	36	0	0	0	0	36	0	0	0	0	34	2	0
	6 to 2	0	2	34	0	0	0	0	36	0	0	1	0	35	0	0
	6 to 5	0	0	36	0	0	0	0	36	0	0	0	0	36	0	0
	6 to 7	1	5	26	4	0	0	3	31	2	0	0	2	34	0	0
<b>Communities</b>	1 to 2	0	0	36	0	0	0	0	36	0	0	0	0	36	0	0
	1 to 3	0	2	34	0	0	0	0	36	0	0	0	2	34	0	0
<b>Culture / Ethnicity</b>	5 to 0	0	2	33	1	0	0	0	36	0	0	0	2	34	0	0
	5 to 1	0	0	36	0	0	0	0	36	0	0	0	1	35	0	0
	5 to 2	0	0	36	0	0	0	1	34	1	0	0	3	32	1	0
	5 to 3	0	0	36	0	0	0	1	35	0	0	0	1	35	0	0
	5 to 4	0	0	36	0	0	0	0	36	0	0	0	1	35	0	0
<b>Gender</b>	M to F	0	0	35	1	0	0	0	36	0	0	1	1	34	0	0
<b>Special Programs</b>	1 to 2	0	1	35	0	0	0	0	36	0	0	0	0	35	1	0
	1 to 3	0	1	35	0	0	0	0	35	1	0	0	1	34	1	0

**Table 52 – Summary of Measured DIF – Item Summary Table – Benchmark 2**

Number of items of the type showing the amount of DIF for the column. For DIF Scale, see DIF Rating Criteria chart.																
Categories of Comparisons		Multiple Choice Items and Constructed Response Items Separately Compared														
		Number of Items with each value on the DIF Scale														
		Math Items = 36					Reading Items = 36					Writing Items = 36				
DIF Scale =		2	1	0	-1	-2	2	1	0	-1	-2	2	1	0	-1	-2
<b>Regions</b>	6 to 1	0	0	36	0	0	0	0	36	0	0	0	0	36	0	0
	Interior.....1	0	1	31	3	1	0	1	34	1	0	0	2	32	2	0
	North Arctic.....2	0	1	35	0	0	0	0	36	0	0	0	0	36	0	0
	Southeast.....5	0	2	32	1	1	0	0	35	1	0	1	3	27	5	0
	Southern.....6															
	Yukon/Kus.....7															
<b>Communities</b>	1 to 2	0	1	35	0	0	0	0	36	0	0	0	0	36	0	0
	Urban.....1															
	Rural.....2															
	1 to 3	0	1	32	2	1	0	0	36	0	0	0	2	34	0	0
	Remote.....3															
<b>Culture / Ethnicity</b>	5 to 0	0	0	34	1	1	0	1	35	0	0	0	3	30	3	0
	Alaskan Native...0															
	5 to 1	0	0	35	1	0	0	0	36	0	0	0	0	36	0	0
	Native Amer.....1															
	5 to 2	0	2	33	0	1	0	2	34	0	0	1	3	29	1	2
	Asian.....2															
	5 to 3	0	0	36	0	0	0	1	35	0	0	0	1	35	0	0
	Afr. Amer.....3															
	5 to 4	0	0	35	1	0	0	0	36	0	0	0	1	35	0	0
	Hispanic Amer...4															
	Caucasian.....5															
<b>Gender</b>	M to F	1	0	34	1	0	0	1	35	0	0	1	4	30	1	0
	Male and Female															
<b>Special Programs</b>	1 to 2	0	0	36	0	0	0	0	36	0	0	0	0	35	1	0
	Regular.....1															
	1 to 3	0	0	35	0	1	0	1	34	1	0	0	4	29	3	0
	Special Edu.....2															
	LEP.....3															

**Table 53 – Summary of Measured DIF – Item Summary Table – Benchmark 3**

<b>Number of items of the type showing the amount of DIF for the column.</b> <b>For DIF Scale, see DIF Rating Criteria chart.</b>																
<b>Categories of Comparisons</b>		<b>Multiple Choice Items and Constructed Response Items Separately Compared</b>														
		<b>Number of Items with each value on the DIF Scale</b>														
		<b>Math Items = 36</b>					<b>Reading Items = 36</b>					<b>Writing Items = 36</b>				
<b>DIF Scale =</b>		<b>2</b>	<b>1</b>	<b>0</b>	<b>-1</b>	<b>-2</b>	<b>2</b>	<b>1</b>	<b>0</b>	<b>-1</b>	<b>-2</b>	<b>2</b>	<b>1</b>	<b>0</b>	<b>-1</b>	<b>-2</b>
<b>Regions</b>	6 to 1	0	0	36	0	0	0	0	36	0	0	0	0	34	2	0
	6 to 2	1	1	32	2	0	0	2	33	1	0	0	2	32	2	0
	6 to 5	0	0	36	0	0	0	0	36	0	0	0	0	36	0	0
	6 to 7	0	2	33	1	0	0	3	30	3	0	0	2	31	3	0
	Yukon/Kus.....7															
<b>Communities</b>	1 to 2	0	0	36	0	0	0	0	36	0	0	0	0	36	0	0
	Urban.....1															
	1 to 3	0	1	33	2	0	0	2	34	0	0	0	1	33	2	0
	Rural.....2															
	Remote.....3															
<b>Culture / Ethnicity</b>	5 to 0	0	1	35	0	0	0	2	34	0	0	0	2	32	2	0
	Alaskan Native...0															
	5 to 1	0	1	33	2	0	0	1	35	0	0	0	0	36	0	0
	Native Amer.....1															
	5 to 2	0	1	35	0	0	0	0	36	0	0	1	4	29	2	0
	Asian.....2															
	5 to 3	0	1	35	0	0	0	0	35	1	0	0	4	32	0	0
	Afr. Amer.....3															
	5 to 4	0	0	36	0	0	0	0	36	0	0	0	2	33	1	0
	Hispanic Amer...4															
	Caucasian.....5															
<b>Gender</b>	M to F	0	0	36	0	0	0	1	35	0	0	1	4	31	0	0
	Male and Female															
<b>Special Programs</b>	1 to 2	0	0	36	0	0	0	0	36	0	0	0	0	35	1	0
	Regular.....1															
	1 to 3	0	1	34	1	0	0	2	34	0	0	0	5	28	3	0
	Special Edu.....2															
	LEP.....3															

## Alaska Performance Index

Mastery of the objective for other assessments has previously been reported in terms of the percent of points a student obtained out of the total points possible. Specifically, a student was said to have mastered a content performance standard if the student obtained 75% or more correct out of the total possible points on a given objective (content performance standard).

This definition of mastery tended to produce unstable results from year to year for two reasons. First, the ability to achieve mastery is highly dependent on the difficulty of the items contributing to a given content performance standard. If somewhat easier items are used to measure a content performance standard in one year than in the previous year, a greater percentage of students will achieve mastery in the year with the easier items, even if the students are of equal ability in both years. Second, each content performance standard is measured by a relatively small number of items—some content performance standards are measured by as few as four score points. In general, longer measures produce more reliable results and shorter measures produce less stable results.

Although the overall difficulty of the test is controlled from year to year, it is not controlled at the content performance standard level. To accomplish this, extremely large numbers of items would have to be written and piloted in order to select items with the same difficulty as those from the previous year. This would be extremely expensive, and for all practical purposes, is not an option.

Several options were considered to achieve stability at the content performance standard level. First, because standard equating procedures are used to account for differences in overall test difficulty from year to year, one might consider using these procedures at the content performance standard level. That is, produce a scale score associated with each content performance standard, just as a scale score is produced at the total test level. If the equating procedures *were* applied, however, the relatively small number of score points contributing to each content performance standard would result in equated scale scores with very large standard errors, and hence, the results would remain unstable.

A two-prong solution to the problem was considered next. First, to help ameliorate the instability resulting from the relatively small number of points contributing to each content performance standard, the Alaska Performance Index (API) was proposed. The API uses Bayesian statistics to add stability to the individual student results for each content performance standard. The API expresses a student's content performance standard score as an estimate of the number of items the student would respond to correctly out of 100 items similar to those actually used to measure the content performance standard. The API has been demonstrated to be more reliable than simply reporting the number or percent correct. A more thorough description of the computation of the API procedures can be found in Appendix B.

Second, to control for differences in the difficulty of the content performance standard items from year to year, a stable reference point was needed. The student at the borderline of the standard cut-score provided a convenient and relevant reference point. For each content performance standard, the API was calculated so that it would reflect what would be expected for a student exactly at the cut score that defines the standard for the given content area. This reference point changes, as desired, with item difficulty. If more difficult items contribute to a content performance standard, the student at the standard would be expected to achieve a lower

API than if less difficult items were used. The API cut scores for the Benchmark and HSGQE Spring 2002 operational forms are tabulated in *The Standard Setting Technical Reports for Benchmark* and *HSGQE*, respectively.

### API Cutpoint

Tables 54 – 56 present the cutpoints for the four (4) levels of proficiency set by the Benchmark standard setting. The tables present the levels of proficiency by content area and grade level by individual objective.

**Table 54 – Benchmark 1 API Cutpoints**

<b>Benchmark 1 Alaska Performance Index</b>					
Grade 3 Content Areas		Not Prof.	Below Prof.	Prof.	Adv.
Math Objective Codes	M.A.1	0.00 – 0.36	0.37 – 0.57	0.58 – 0.76	0.77 – 0.99
	M.A.2	0.00 – 0.41	0.42 – 0.61	0.62 – 0.82	0.83 – 0.99
	M.A.3	0.00 – 0.36	0.37 – 0.57	0.58 – 0.81	0.82 – 0.99
	M.A.4	0.00 – 0.48	0.49 – 0.66	0.67 – 0.84	0.85 – 0.99
	M.A.5	0.00 – 0.31	0.32 – 0.49	0.50 – 0.69	0.70 – 0.99
	M.A.6	0.00 – 0.53	0.54 – 0.67	0.68 – 0.79	0.80 – 0.99
	B/C/D	0.00 – 0.49	0.50 – 0.66	0.67 – 0.79	0.80 – 0.99
Reading Objective Codes	R.01	0.00 – 0.62	0.63 – 0.76	0.77 – 0.92	0.93 – 0.99
	R.02	0.00 – 0.42	0.43 – 0.61	0.62 – 0.88	0.89 – 0.99
	R.04	0.00 – 0.41	0.42 – 0.54	0.55 – 0.74	0.75 – 0.99
	R.05	0.00 – 0.42	0.43 – 0.59	0.60 – 0.78	0.79 – 0.99
	R.06	0.00 – 0.27	0.28 – 0.38	0.39 – 0.75	0.76 – 0.99
	R.07	0.00 – 0.30	0.31 – 0.44	0.45 – 0.77	0.78 – 0.99
	R.08	0.00 – 0.47	0.48 – 0.66	0.67 – 0.82	0.83 – 0.99
	R.09	0.00 – 0.59	0.60 – 0.79	0.80 – 0.97	0.98 – 0.99
Writing Objective Codes	w1/2	0.00 – 0.40	0.41 – 0.62	0.63 – 0.78	0.79 – 0.99
	w3	0.00 – 0.20	0.21 – 0.46	0.47 – 0.84	0.85 – 0.99
	w4	0.00 – 0.24	0.25 – 0.45	0.46 – 0.79	0.80 – 0.99

**Table 55 – Benchmark 2 API Cutpoints**

<b>Benchmark 2 Alaska Performance Index</b>					
Grade 6 Content Areas		Not Prof.	Below Prof.	Prof.	Adv.
Math Objective Codes	M.A.1	0.00 – 0.54	0.55 – 0.69	0.70 – 0.88	0.89 – 0.99
	M.A.2	0.00 – 0.52	0.53 – 0.67	0.68 – 0.88	0.89 – 0.99
	M.A.3	0.00 – 0.50	0.51 – 0.62	0.63 – 0.81	0.82 – 0.99
	M.A.4	0.00 – 0.44	0.45 – 0.52	0.53 – 0.66	0.67 – 0.99
	M.A.5	0.00 – 0.25	0.26 – 0.34	0.35 – 0.54	0.55 – 0.99
	M.A.6	0.00 – 0.27	0.28 – 0.41	0.42 – 0.68	0.69 – 0.99
	B/C/D	0.00 – 0.16	0.17 – 0.27	0.28 – 0.51	0.52 – 0.99
Reading Objective Codes	R.01	0.00 – 0.38	0.39 – 0.59	0.60 – 0.78	0.79 – 0.99
	R.02	0.00 – 0.40	0.41 – 0.62	0.63 – 0.78	0.79 – 0.99
	R.04	0.00 – 0.58	0.59 – 0.74	0.75 – 0.84	0.85 – 0.99
	R.05	0.00 – 0.25	0.26 – 0.38	0.39 – 0.51	0.52 – 0.99
	R.06	0.00 – 0.46	0.47 – 0.63	0.64 – 0.72	0.73 – 0.99
	R.07	0.00 – 0.30	0.31 – 0.48	0.49 – 0.70	0.71 – 0.99
	R.08	0.00 – 0.44	0.45 – 0.68	0.69 – 0.82	0.83 – 0.99
	R.09	0.00 – 0.28	0.29 – 0.43	0.44 – 0.62	0.63 – 0.99
Writing Objective Codes	w1/2	0.00 – 0.27	0.28 – 0.46	0.47 – 0.65	0.66 – 0.99
	w3	0.00 – 0.27	0.28 – 0.53	0.54 – 0.79	0.80 – 0.99
	w4	0.00 – 0.27	0.28 – 0.53	0.54 – 0.84	0.85 – 0.99



**Table 56 – Benchmark 3 API Cutpoints**

<b>Benchmark 3 Alaska Performance Index</b>					
Grade 8 Content Areas		Not Prof.	Below Prof.	Prof.	Adv.
Math Objective Codes	M.A.1	0.00 – 0.33	0.34 – 0.69	0.70 – 0.84	0.85 – 0.99
	M.A.2	0.00 – 0.31	0.32 – 0.51	0.52 – 0.76	0.77 – 0.99
	M.A.3	0.00 – 0.36	0.37 – 0.63	0.64 – 0.89	0.90 – 0.99
	M.A.4	0.00 – 0.39	0.40 – 0.74	0.75 – 0.88	0.89 – 0.99
	M.A.5	0.00 – 0.55	0.56 – 0.79	0.80 – 0.93	0.94 – 0.99
	M.A.6	0.00 – 0.29	0.30 – 0.51	0.52 – 0.74	0.75 – 0.99
	B/C/D	0.00 – 0.25	0.26 – 0.45	0.46 – 0.99	N/A
Reading Objective Codes	R.01	0.00 – 0.41	0.42 – 0.52	0.53 – 0.66	0.67 – 0.99
	R.10	0.00 – 0.33	0.34 – 0.43	0.44 – 0.59	0.60 – 0.99
	R.04	0.00 – 0.46	0.47 – 0.59	0.60 – 0.74	0.75 – 0.99
	R.05	0.00 – 0.46	0.47 – 0.58	0.59 – 0.76	0.77 – 0.99
	R.06	0.00 – 0.60	0.61 – 0.73	0.74 – 0.85	0.86 – 0.99
	R.07	0.00 – 0.30	0.31 – 0.39	0.40 – 0.55	0.56 – 0.99
	R.08	0.00 – 0.44	0.45 – 0.54	0.55 – 0.67	0.68 – 0.99
	R.09	0.00 – 0.53	0.54 – 0.64	0.65 – 0.79	0.80 – 0.99
Writing Objective Codes	w1/2	0.00 – 0.23	0.24 – 0.53	0.54 – 0.69	0.70 – 0.99
	w3	0.00 – 0.29	0.30 – 0.58	0.59 – 0.79	0.80 – 0.99
	w4	0.00 – 0.28	0.29 – 0.60	0.61 – 0.83	0.84 – 0.99

Grade 8 Mathematics Objective B/C/D Advanced Proficient criteria is no longer posted due to the elimination of item addressing the objective.

The API cutpoints for the HSGQE is shown on Table 57.

**Table 57 – HSGQE API Cutpoints**

Content Area	Objective Code	Title	Not Proficient	Proficient
Reading	R4.1	Use context clues	0.00 – 0.62	0.63 – 0.99
	R4.4	Summarize information	0.00 – 0.68	0.69 – 0.99
	R4.5	Critique arguments	0.00 – 0.70	0.71 – 0.99
	R4.6	Apply multi-step directions	0.00 – 0.69	0.70 – 0.99
	R4.9	Make and support assertions	0.00 – 0.54	0.55 – 0.99
	R4.10	Analyze and evaluate themes	0.00 – 0.59	0.60 – 0.99
Math	A1	Numeration	0.00 – 0.61	0.62 – 0.99
	A2	Measurement	0.00 – 0.52	0.53 – 0.99
	A3	Estimation & Computation	0.00 – 0.48	0.49 – 0.99
	A4	Functions & Relationships	0.00 – 0.50	0.51 – 0.99
	A5	Geometry	0.00 – 0.42	0.43 – 0.99
	A6	Statistics/Probability	0.00 – 0.56	0.57 – 0.99
	BCD	Prob. Solve/Comm/Reasoning	0.00 – 0.24	0.25 – 0.99
Writing	W4.1/4.2	Write compositions	0.00 – 0.51	0.52 – 0.99
	W4.3	Use conventional English	0.00 – 0.60	0.61 – 0.99
	W4.4	Revise writing for word choice	0.00 – 0.60	0.61 – 0.99

## HSGQE Field Test

This section contains information relating to the items field tested in the Spring 2003 assessment. In order to field test enough items to continue to improve the item pool needed for future assessments and also keep the individual testing time to a minimum nine forms were used. The operational items are given first followed by the field test items. Field test items do not contribute to the individual's total score. The following tables 58 – 61 summarize the field test (FT) item information. This includes number of FT items by item type, item analysis, summary of calibration results, and item fit and non-convergence.

**Table 58 – Number of Field Test Items Administered by Item Type for HSGQE**

Content Area	Form	SCORE POINTS							Total CR Points
		MC	1 pt SCR	2 pt SCR	3 pt SCR	4 pt ECR	6 pt ECR	Total	
Reading	FT1	12		3				15	6
	FT2	10		3	2			15	12
	FT3	10		4	1			15	11
	FT4	10		3	2			15	12
	FT5	11		3	1			15	9
	FT6	11		1	3			15	11
	FT7	12		1	1	1		15	9
	FT8	10		2	3			15	13
	FT9	10		2	2	1		15	14
Mathematics	FT1	11		1		1		13	6
	FT2	11		1		1		13	6
	FT3	11		1		1		13	6
	FT4	11		1		1		13	6
	FT5	11		1		1		13	6
	FT6	11		1		1		13	6
	FT7	10		3				13	6
	FT8	10		3				13	6
	FT9	10		3				13	6
Writing	FT1						1	1	6
	FT2						1	1	6
	FT3						1	1	6
	FT4	1	1	2		2		6	13
	FT5	1	1	2		2		6	13
	FT6	1		2		3		6	16
	FT7	1		2		3		6	16
	FT8	7	1	1		2		11	11
	FT9	7	1	1		2		11	11

**Table 59 – Field Test P – Value Results**

Content Area	Form	P – Value Mean	Item Difficulty ( P – Value )	
			Lowest = Most difficult Item	Highest = Easiest Item
Reading	FT1	0.63	0.23	0.82
	FT2	0.53	0.21	0.86
	FT3	0.67	0.38	0.87
	FT4	0.69	0.57	0.87
	FT5	0.71	0.43	0.94
	FT6	0.72	0.60	0.86
	FT7	0.68	0.45	0.90
	FT8	0.69	0.30	0.88
	FT9	0.79	0.45	0.92
Mathematics	FT1	0.62	0.36	0.87
	FT2	0.56	0.03	0.84
	FT3	0.65	0.19	0.89
	FT4	0.60	0.37	0.95
	FT5	0.58	0.18	0.86
	FT6	0.54	0.25	0.71
	FT7	0.46	0.10	0.80
	FT8	0.63	0.30	0.92
	FT9	0.53	0.18	0.81
Writing	FT1	0.56	0.56	0.56
	FT2	0.42	0.42	0.42
	FT3	0.54	0.54	0.54
	FT4	0.64	0.56	0.73
	FT5	0.63	0.57	0.69
	FT6	0.56	0.38	0.70
	FT7	0.66	0.56	0.78
	FT8	0.59	0.18	0.87
	FT9	0.68	0.56	0.88

**Table 60 – Summary of Calibration Results on Operational and Field Test Items – HSGQE**

Content Area	Total Number of Items (across all forms)	Max A	Default C	B Value Range (scale score matrix)	No. of Est. Cycles	Non – Conv Items
RD	185	0	41	-4.404 to 3.730	50	2
MA	175	1	33	-4.002 to 4.373	50	2
WR	84	0	21	-3.460 to 3.840	50	0

**Table 61 – Number of Misfit Field Test Items – HSGQE**

Content Area	Total # of Items	Number of Misfit Items	Percentage of Misfit Items
RD	135	20	15
MA	117	8	7
WR	49	23	47

## Appendix A: Fit Measurement: A Generalization of Q1

Referring to Yen (1981, p. 246–247), it can be seen that Q1 is a Pearson chi-square of the form

$$\chi^2 = \sum_{i=1}^I \frac{(f_{oi} - f_{ei})^2}{f_{ei}},$$

where  $f_{oi}$  and  $f_{ei}$  are observed and expected frequencies of observations failing in cell  $i$ . In using the Pearson  $\chi^2$ , every sample observation must fall in one and only one cell, the observation must be independent, and  $N$  large. To get  $Q_{ij}$  for binary items

$$Q_{1j} = \sum_{i=1}^I \frac{N_{ji}(O_{ji} - E_{ji})^2}{E_{ji}} + \sum_{i=1}^I \frac{N_{ji}[(1 - O_{ji}) - (1 - E_{ji})]^2}{E_{ji}} \quad (1)$$

$$= \sum_{i=1}^I \frac{N_{ji}(O_{ji} - E_{ji})^2}{E_{ji}(1 - E_{ji})}. \quad (2)$$

$N_{ji}$  is the number of examinees in cell  $i$  for item  $j$ ;  $O_{ji}$  and  $E_{ji}$  are the observed and predicted proportions of examinees in cell  $i$  that pass item  $j$ :

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a \in I}^{N_{ji}} P_j(\hat{\theta}_a). \quad (3)$$

[It can be noted (see Yen, 1981) that for dichotomous items, Q1 measures fit essentially the same way as the fit measure used in BICAL by Wright and Mead (1977).]

The generalization of Q1 for items with multiple response categories can be stated as

$$Q_{1j} = \sum_{i=1}^I \sum_{k=1}^{m_j} \frac{N_{jki}(O_{jki} - E_{jki})^2}{E_{jki}} \quad (4)$$

with

$$E_{jki} = \frac{1}{N_{ji}} \sum_{a \in I}^{N_{ji}} P_{jk}(\hat{\theta}_a). \quad (5)$$

$O_{jki}$  is the observed proportion of examinees in cell  $i$  who are at level  $k$ . It can be noted that (4) is equivalent to the description of Q1 in (1), making it straightforward to obtain fit when 3PL (or 1PL) and 2PPC (1PPC) items both occur.

The degrees of freedom for  $Q_{Ij}$  are the number of independent observations entering into the calculation minus the number of independent parameters estimated for that item. It is known that

$$\sum_{k=1}^{m_j} O_{jki} = 1,$$

meaning that there are  $m_j - 1$  independent observations per cell, giving a total of  $I(m_j - 1)$  independent observations for item  $j$ . For the 3PL, the number of independent parameters estimated is 3. For the 2PPC, there are  $m_j$  independent parameters ( $\alpha_j$  and the  $m_j - 1$  values of  $o_{ji}$ ). For the Partial Credit model, there are  $m_j - 1$  values of  $o_{ji}$  estimated per item. Thus,  $Q_{Ij}$  is assumed to have approximately a chi-square distribution with the following degrees of freedom:

$$3\text{PL: } df = I \cdot (m_j - 1) - 3 = 7$$

$$1\text{PL: } df = I \cdot (m_j - 1) - 1 = 9$$

$$2\text{PPC: } df = I \cdot (m_j - 1) - m_j$$

$$1\text{PPC: } df = I \cdot (m_j - 1) - (m_j - 1)$$

Given these differences in degree of freedom, it is awkward to compare chi-square values across items. Therefore, a standardization is produced:

$$Z_{Q_{Ij}} \equiv \frac{Q_{Ij} - df}{(2df)^{1/2}}. \quad (6)$$

## Appendix B: Alaska Performance Index Procedure

An Alaskan Performance Index (API) is reported for content performance standards in the Benchmark and the HSGQE assessments; a confidence interval, which reflects the standard error of measurement of the API, is also reported for each API for an individual. Because there are small numbers of items in many content performance standards, the confidence intervals can be very wide if they are based only on the information contained in each content performance standard. In order to decrease the standard error of the API, information from the examinee's related performance on the rest of the items in the test is taken into account in calculating the API; the incorporation of the information produces narrower confidence intervals. A summary of the procedure is provided below.<sup>2</sup>

### Summary of the API Procedure

Step 1. Estimate IRT parameters  $a_i$ ,  $b_i$ , and  $c_j$  from the operational data for each assessment.

Step 2. Treating the item parameter estimates as fixed values, obtain  $\hat{\theta}$  for each examinee based on overall test performance.

Step 3. For each content performance standard calculate

$$P_i(\hat{\theta}) = c_{ij} + \frac{1 - c_{ij}}{1 + \exp\left[-1.7a_{ij}\left(\hat{\theta} - b_{ij}\right)\right]}$$

and

$$\hat{T}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\hat{\theta}).$$

Step 4. Obtain

$$Q = \sum_{j=1}^J \frac{n_j \left( \frac{x_j}{n_j} - \hat{T}_j \right)^2}{\hat{T}_j \left( 1 - \hat{T}_j \right)}.$$

Step 5. If  $Q \leq X^2(J, .10)$ ,

$$p_j = \hat{T}_j n_j^* + X_j$$

and

---

<sup>2</sup> CAT/5 Technical Bulletin 2 provides a detailed explanation of the derivation of the combination Bayesian/Item Response Theory (IRT) procedure used for calculating the API and its confidence interval. The assumptions used in the derivations are also discussed in the Technical Bulletin. Empirical evidence supporting the accuracy of the confidence interval can be found in Yen (1987).



$$q_j = \left(1 - \hat{T}_j\right) n_j^* + n_j - x_j.$$

The API is defined to be

$$\begin{aligned}\tilde{T}_j &= \frac{p_j}{p_j + q_j} \\ &= \frac{\hat{T}_j n_j^* + x_j}{n_j^* + n_j}.\end{aligned}$$

The value of  $I(\theta, \hat{\theta})$  used in calculating  $\sigma^2(\hat{T}_j | \theta)$  and  $n_j^*$  is based on

$$I(\theta, \hat{\theta}) = \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{[P_{ij}(\theta)]^2}{[P_{ij}(\theta)][1 - P_{ij}(\theta)]} \text{ for item pattern scoring}$$

or

$$I(\theta, \hat{\theta}) = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} [P_{ij}(\theta)]^2}{\sum_{j=1}^J \sum_{i=1}^{n_j} [P_{ij}(\theta)][1 - P_{ij}(\theta)]} \text{ for number correct scoring}$$

method to obtain  $\hat{\theta}$ . The values of  $p_j$  and  $q_j$  are used to obtain the 67% credibility interval for  $T_j$  from the beta distribution.

If  $Q \leq X^2(J, .10)$ ,

$$\begin{aligned}\tilde{T}_j &= x_j / n_j, \\ p_j &= X_j,\end{aligned}$$

and

$$q_j = n_j - X_j.$$

## References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37: 29–51.

Brown, F. G. (1976). *Principles of Educational and Psychological Testing*, 2nd ed. New York: Holt, Rinehart, and Winston.

Burket, G. R. (1988). ITEMSYS [Computer Program]. Monterey, CA: CTB/McGraw-Hill.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Divgi, D. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23: 283–298.

Feldt, L. S., and Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.), *Educational Measurement, Third Edition*, 105–146. Washington, D.C.: American Council on Education.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education*, 2: 297–312.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 18: 109–118.

Linn, R. L., Levine, M. V., Hasting, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 18: 109–118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Macmillan/McGraw-Hill. (1993). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*. New York: Author.

McGraw-Hill. (1983). *Guidelines for bias-free publishing*. Monterey, CA: Author.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and*

*Psychology*, 3rd ed. New York: Holt, Rinehart, and Winston.

Sandoval, J. H., & Mille, M. P. (1979, August). Accuracy of judgments of WISC-R item difficulty for minority groups. Paper presented at the annual meeting of the American Psychological Association, New York.

Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist*, 19: 219–225.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7: 201–210.

Thissen, D., Wainer, H., & Wang, X-B. (1992). How unidimensional are tests comprising both multiple-choice and free-response items? An analysis of two tests, ETS Technical Report 92 Princeton, NJ: Educational Testing Service.

Wright, B. D., & Mead, R. J. (1977). *BICAL: Calibrating items and scales with the Rasch Model*. (Research Memorandum No. 23). Chicago, IL: Department of Education, Statistical Laboratory.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5: 245–262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, 21: 93–111.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30: 233–251.